Università Politecnica delle Marche

Facoltà di Ingegneria

Dipartimento di Ingegneria dell'Informazione

Dottorato di ricerca in ingegneria dell'informazione,
curriculum informatica, gestionale e dell'automazione

Open Data Analytics

Advanced methods, tools and visualizations for policy making

Supervisor(s):

Emanuele Frontoni

Donato Iacobucci

Discussant(s):

Alessandro Aldini

Ugo Fratesi

Phd candidate:

Roberto Palloni

Triennio accademico 2015/2018 - 17° ciclo

A Eleonora,
complice e sostegno

Ad Andrea,
sempre con noi

Open Data Analytics
Advanced methods, tools and visualizations for policy making

Table of contents

List of figures

List of tables

List of acronyms

Acronym	**Full term**
AIR	Annual Implementation Report
API	Application program interface
ARI	Average Relatedness Index
BI	Business Intelligence
CPR	Common Provisions Regulation
CORDIS	Community Research and Development Information Service
DBMS	Database Management System
EAFRD	European Agricultural Fund for Rural Development
EC	European Commission
ELT	Extract, Load, Transform (operations)
EMFF	European Maritime and Fishery Fund
ERDF	European Regional Development Fund
ESIF	European Structural and Investment Funds
ESF	European Social Fund
ESPON	European Spatial Planning Observation Network
ETC	European Territorial Cooperation
ETL	Extraction Transformation and Loading
EU	European Union
FP7	7th Framework Programme for Research and Technological Development
KPI	Key Performance Indicators
IPC	International Patent Classification
MA	Managing authority
NACE	Nomenclature statistique des activités économiques
NPL	Natural Language Processing
NUTS	Nomenclature of Territorial Units for Statistics

OP	Operational Programme
OLAP	Online analytical processing
PA	Priority Axis
PaaS	Platform as a Service
REST	Representational State Transfer
R&D	Research and development
RCA	Relative Comparative Advantage
RSI	Related Share Index
S3	Smart Specialisation Strategy
SFC	EU system for fund data management
TO	Thematic Objective
WIPO	World Intellectual Property Organization

Foreword

This document is the final version of the PhD thesis defended by Roberto Palloni as the final outcome of the three years (2015-2018) PhD programme at Università Politecnica delle Marche – Department of Information Engineering.

This thesis is the result of research under the supervision of two academic tutors, Professor Emanuele Frontoni and Professor Donato Iacobucci as well as Andrea Gramillano, t33 senior policy analyst.

The document has been reviewed and assessed by Professor Ugo Fratesi (Politecnico di Milano) and Professor Alessandro Aldini (Università degli studi di Urbino).

The thesis is structured into five chapters combining two principal areas of research under the common topic of public policy open data.

Chapter 1 is an introduction to the concept of open data and the transformation of public administration in progress as more and more decisions rely on data science.

Chapter 2 describes the status of data production and use at EU level, in particular for national and regional public investments financed by European Structural and Investment Funds (ESIF).

In response to literature critical of the abstract planning processes based on weak evidences and the adoption of open data initiatives as mere repositories of data, chapter 3 introduces the topic of open data analytics and visualization.

This includes examples of tools and methods currently available to support and ease policy decision using open data. This chapter also presents the current limits and partial use of data by these tools to introduce chapter 4 which describes ESIFy, the tool developed as a side project of this thesis.

This web application allows to explore ESIF open data to easily visualize the implementation and performance of investments across Europe as an attempt to improve the current state of the art.

Chapter 5 further extends the use of open data sources beyond monitoring and performance assessment to being able to support strategy planning. In particular, the chapter focuses on decision-making support for European innovation policy and the allocation of crucial regional investments in public and private research and development (R&D). A final consideration chapter concludes the research with a look into the future of the next programming period 2021-2027 and the potential for a new mindset of *smart* data.

Summary

The research discussed in this thesis is focused on developing and applying new methodologies for collecting, processing and visualizing large sets of open data for public policy performance assessment and decision making.

The research focuses on the effectiveness of ESIF and the use of other open data sources for data-driven decision making supporting public managers. While data analytics represents the research **topic**, public policy open data is the application **domain**.

Beyond the problem of transparency and accountability to citizens, accessible and usable public data have great informative value for public administration decision making. Open sources of data allow administrators, researchers and practitioners to develop new analysis and visualizations that can unleash the hidden informative potential of data with meaningful insights.

Data management and performance monitoring are core to business intelligence (BI) informing and supporting decision making, not only for private enterprises.

For this reason, recent years have seen increasing numbers of open data initiatives, public database diffusion, open data hackathons and data-related initiatives[1].

Furthermore, the *pari passu* diffusion of recent technologies and data sharing systems such as application program interfaces (APIs) are also boosting the diffusion and use of public policy data.

Based on this framework, using opendata the research attempts to address the following two hypotheses:

> ***H1****: Open data platforms can be considered useful for policy making and not just as data tombs set up only to satisfy governmental digital agenda requirements.*
>
> ***H2:*** *In allocating research and innovation investments, regions have developed their S3s according to embeddedness, relatedness and connectivity.*

For this reason, this document has two main parts:

- The first gives a comprehensive overview of the use of open data at European Union (EU) level for monitoring and performance assessment. Using ESIF open data, the research focuses on developing a wider and deeper approach to the use of open data for simpler and more effective interpretation and insights;

[1] The Global Open Data Index provides an overview of the state of open government data publication https://index.okfn.org/

- The second part uses additional open data sources (public documents, RegPat patents and CORDIS[2] projects) to assist strategy development and assessment. The research in this second part highlights possible uses of open data to promote data-driven policy making.

In the first part is an explanation of a web tool dedicated to visualizing ESIF open data that improves on current methodologies and tools to expose insights based on:

- ✓ Theoretical principles for simpler and more effective interpretation;
- ✓ Adoption of advanced technologies for a simple and flexible solution.

The second part includes an analytical framework with empirical results. This is based on the capacity of Italian regions to effectively allocate crucial regional investments in public and private R&D, according to the European innovation policy (Smart Specialisation Strategy).

Despite the wider diffusion and increase of open data availability and of more powerful technologies to exploit their potential, these resources present many issues (Schintler, 2014). Beyond the volume and the complexity of the available data, the velocity and veracity (uncertainty) of the data sources affect their quality, accuracy and completeness. ESIF open public information still lacks sufficient granularity and completeness to represent a full informative asset. This creates a blurred lens problem, with reduced informative power for the data available. For example, microdata on individual projects financed by ESIF are reported by only a few Member States despite they are the most detailed source of information at the deepest level possible for the problem under analysis.

Moreover, using open data often implies adopting only an approximation of the information needed. As with regional investments for innovation policies, patent and research project data are used as a proxy for innovation potential, though most enterprise innovation remains untracked, especially in many Italian regions. Different and more specific data should be collected and assessed to improve decision making regarding such topics.

This problem goes beyond the scope of this document.

[2] CORDIS is the European Commission's primary public repository and portal to disseminate information on all EU-funded research projects and their results in the broadest sense: https://cordis.europa.eu/home_en.html.

1 Introduction

In recent years the political context as well as the financial and economic pressure on EU Member State budgets have imposed a change in terms of accountability and justification for public expenditure. There is a stronger need to inform taxpayers about public investments and to broaden the debate on European economic policy and its future orientation.

At the same time, entities involved in the management of public funds have seen a change in the need for strategic and operative information for decision about how to apportion and optimize public resources allocation.

Recent studies on policy data-related needs identify infrastructure planning (i.e. mobility and transport), regional economic development, and land-use planning as those policy areas that could benefit the most from the use of big data[3] (ESPON, 2018) but, at the same time, are those that are still mostly leaning on more traditional data sources (e.g. statistical office).

Financial constraints have also exacerbated the focus on result-orientation and the performance of authorities in achieving investment objectives defined during design phases. These objectives should be tangible measures of benefits for citizens.

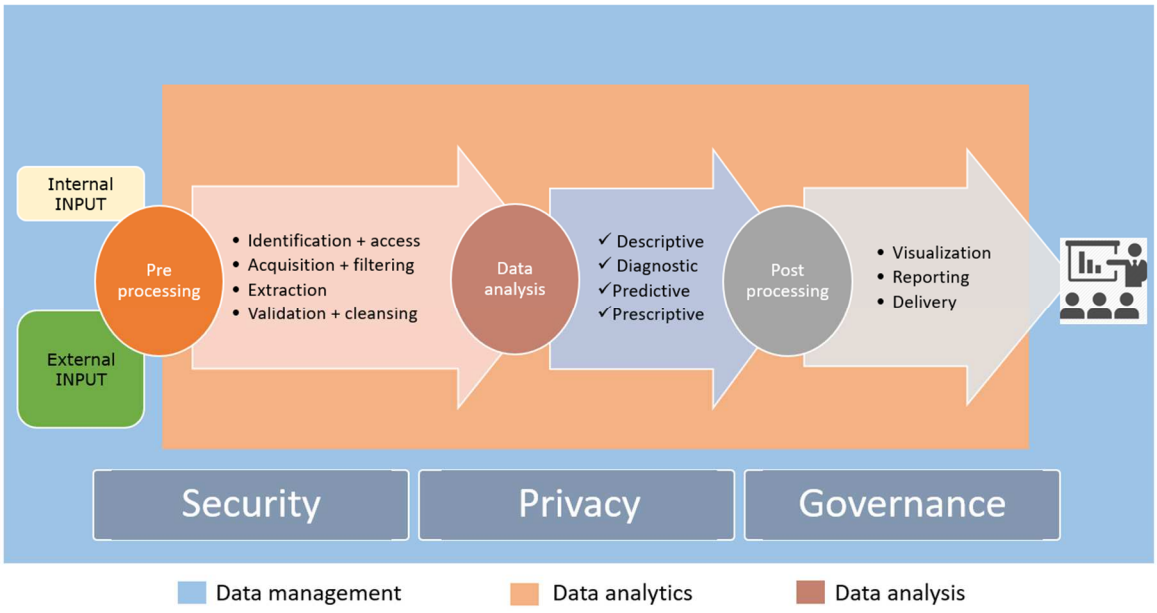
Furthermore, compared to the past, information management capacity has improved significantly, and more quality data is now available due to agile solutions for data sharing and usability.

These aspects have fueled the diffusion of **data analytics** also within public organizations, with a focus on tracking the 'value for money' of public investments.

Beyond traditional data analysis which looks to find patterns, trends and relationships, the new data life cycle covers all pre- and post-processing operations (Erl, Khattak, & Buhler, 2016). Typical data analytics encompasses identifying, accessing, collecting, cleansing, organizing, merging, storing, analyzing, visualizing and reporting (through static or dynamic outputs) diverse types of data. These activities should be considered in the broader framework of data management including data input processes, and the security, privacy and governance of the data system.

[3] Big data describes broadly the volume and the complexity of the available data, as well as sources of data that are too large for traditional processing systems and thus require new technologies (Fawcett, 2014). In addition to the volume, 'big' refers also the variety, velocity and veracity.

Figure 1 – Data management, data analytics and data analysis



Source: Adapted from Erl et al. (Erl et al., 2016)

This data-driven framework could be the basis for other recent phenomena that evolved almost in parallel with the diffusion of data analytics.

The rapid growth in data sources and analytical tools has implied changes in the ways of policy making and its effects on citizens. The combination of the trend of digitizing administrative data, collecting data through diverse devices and rapid development in data storage has led to the establishment of numerous big and open data initiatives at diverse government scale (Giest, 2017).

In response to the new concepts of Digital Era Governance (DEG), Data Readiness and evidence-based policy making and design (Klievink & Cunningham, 2017), the European Commission big data strategy[4] clearly states that data has become a key asset for the European economy and society similar to the classical economic factors, i.e. capital (K) and labour (L) (see also (McKinsey, 2011)).

Within the strategy, encompassing many data-driven subsectors as the cloud computing, industry digitalization, eHealth, Internet Of Things and Smart cities, large space is dedicated to **open data**[5] as the public sector is one of the most data-intensive sectors. European public bodies at all level hold vast amounts of data, known as public sector information (PSI). The EC defines open data as PSI that can be readily and widely accessible and re-used under non-restrictive conditions.

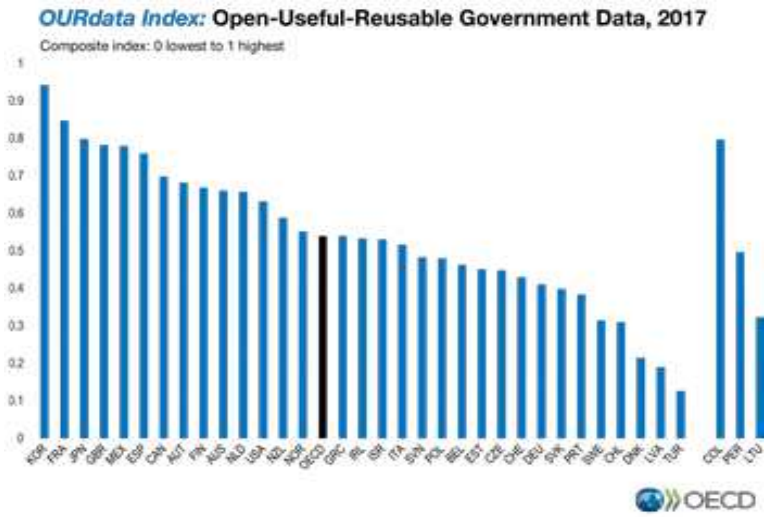
More and more international, national and local public organizations and institutions are releasing quality open data that cover a variety of themes such as the environment,

[4] https://ec.europa.eu/digital-single-market/en/policies/big-data
[5] https://ec.europa.eu/digital-single-market/en/open-data

transport, infrastructure and public fund spending. According to the Open Data maturity report monitoring the status and progress of European countries, the surge of open data is driven by smart cities and in particular by mobility and connection needs (Radu, G. Cecconi, 2018).

The OECD monitors actively the many initiatives worldwide to publish public sector information as open data as presented in the following figure (OECD, 2018).

Figure 2 – Efforts to implement open data in OECD countries



Source: http://www.oecd.org/gov/digital-government/open-government-data.htm

According to the Italian cohesion agency, open data are defined as information published online that is:

- accessible (especially via the Internet) without limitations on to the user's identity or purpose;
- available in a machine language for any application processing without the need for specific software;
- accompanied by metadata and licenses that do not restrict use and re-use.

According to Janssen et al. (Janssen, 2012) open data are produced to be reused in innovative applications. Berners-Lee (Berners-Lee, 2013) and Martin et al. (Martin, Erika G., PhD, MPH; Begany, Grace M., 2018) defined quality open data through the following features:

- Online availability;
- Structured or semi-structured format (e.g. csv, JSON, etc.);
- Usable in free software packages;
- Having a uniform resource identifier;
- Joinable with other data to develop applications.

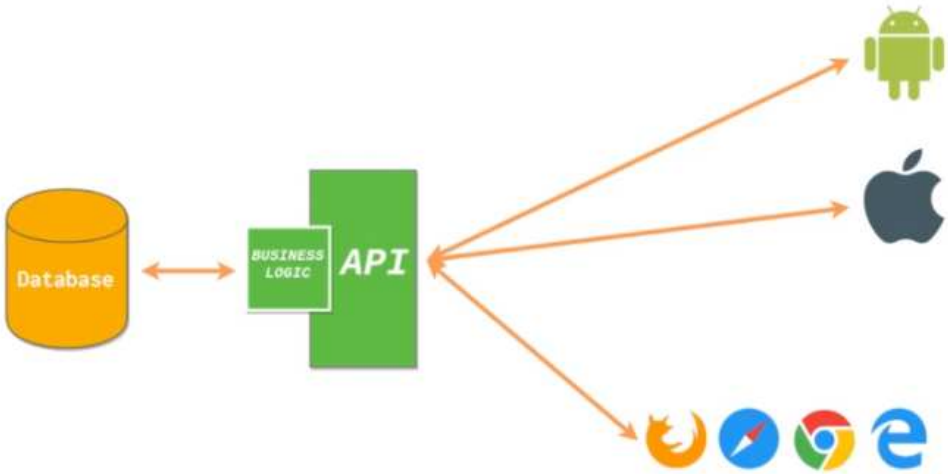
Usually these data are made available through dedicated APIs[6] with REST architecture[7] for easier access and reuse.

The REST acronym stands for Representational State Transfer, which is an architectural design. Usually when we use the term RESTful, we refer to an application implementing the REST architectural design. APIs are the interface part of this architecture and expose and receive data via their **endpoints**.

In other word, an API is the software to interact programmatically at the lower level of the source code, writing functions and algorithms instead of the usual graphical interface.

In short, the main objective of the RESTful architecture is to keep applications back-end and front-end separate in order to easily manage requests from any device and for any purpose.

Figure 3 – Web development with a REST approach



At this regard, in the context of web development, usually when talking about a RESTful API we are referring to Web Services (or Web APIs). Web services are a common way to expose parts of an application to third-parties (external applications and websites). RESTful API usually expose information stored in SQL and NoSQL databases using a common format, such as XML or JSON. This way, any external application can interact with the API, without having to connect directly into the database. Furthermore, it doesn't matter the type of DBMS queried (MySQL, PostgreSQL, MongoDB, etc) or if the application is written in Java, Python or C++ as the API standardize the data flow.

[6] Definition source: https://opencoesione.gov.it/it/open_data/; https://opencoesione.gov.it/it/api-faq/

[7] Representational State Transfer is an architectural style that defines a set of constraints to be used for creating web services. Web services that conform to the REST architectural style, or RESTful web services, provide interoperability between computer systems on the Internet - https://www.w3.org/TR/ws-arch/#relwwwrest.

The framework described implies the presence of **producers** and **consumers** of APIs services and resources. However, this research takes only into account and discusses a real example (ESIFy) of application consuming public APIs endpoints produced by dedicated web services.

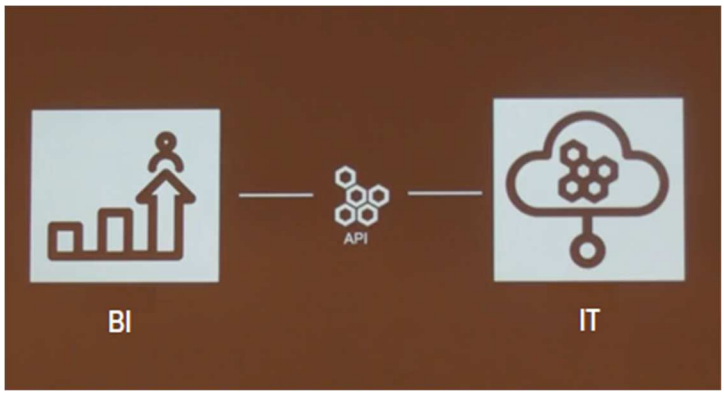
RESTful web services allow requesting systems to access and manipulate data by using a standard and predefined set of operations independently of the programming language and the database system behind them. This implies a larger diffusion and use among developers with different skills and for different purposes. The use of APIs for sharing data, services and business functions between endpoints creates the opportunity to reduce costs and integration time.

Grant (Grant, 2016) provides a non-technical definition of APIs as standardized ways to connect to a database through another interface, and to query the database and get results as data in a standardized manner.

As presented in the following figure, APIs are a modern solution to facilitate data analytics. In the past an IT team was frequently asked to query large databases to produce specific outputs by the analytics team. The newest approach is to allow smaller datasets to be accessed and manipulated in diverse ways directly by a data analyst and data scientists.

Data exposed by the API endpoint could be described as a semi-product that eases the work for both ends of the data processing pipeline. At one end, the analytical process is iterative depending on step-by-step results, where analysts are provided with a dataset instead of a single result. At the other end, the IT team generates a larger set of data without continuously refined queries to the main relational database.

Figure 4 – API as 'back-end for front-end' to align IT and BI



Source: own elaboration

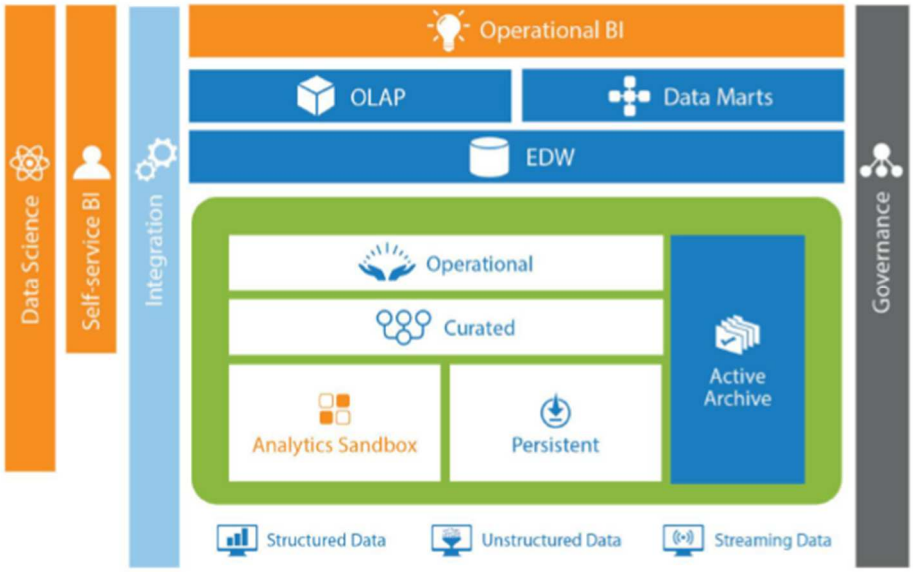
Powerful analytics and visualization services are developed on top of the data layer and making the entire process faster and more flexible.

Furthermore, the increased flexibility of the analytical task is likely to produce parallel results that highlight erroneous or anomalous data, which helps refine data quality and that advise the extraction task with new information.

As data analytics become more and more indispensable to any public and private activity and decision-making process (European Commission, 2017), and both data and sources of data increase significantly, new paradigms are emerging on the IT side for modern data architecture. This involves organizing sources of structured, semi-structured, unstructured and real time data[8] into a persistent data layer and active archive, depending on the immediate or future use.

This data storage area also has an accessible analytics layer for preliminary research and analysis without affecting the persistent strata, i.e. a sandbox. A preliminary analysis enables operational datasets to be organized in the data warehouse layer used by the BI team with online analytical processing (OLAP) and analytical tools. An example of a complex data flow and relative prominent technologies is presented in annex.

Figure 5 – A modern data architecture



Source: D. Ursino, 2018

According to this same paradigm, many public policy open data web services and databases have been made available for reuse in other applications or for analytics. Examples include the Eurostat[9] web service and the OECD[10] web service.

[8] Structured data are organised in relational databases and are estimated to be only 20% of the data; unstructured data, the largest share, has neither a model nor a schema (e.g. text, video, image).
[9] http://ec.europa.eu/eurostat/web/json-and-unicode-web-services
[10] https://data.oecd.org/api/

In public sector, data can increase the efficiency of processes and increase quality and transparency of decision making with substantial cost savings.

Nevertheless, a common criticism about open data is that current efforts are focused on publishing data and not on its **usability**, i.e. how this data is consumed by end users (Helbig, N., Cresswell, A.M., Burke, G.B. and Luna-Reyes, 2012).

Many open data projects have been overly focused on technical issues such as formats, updates and endpoints without caring how this data can be used to produce value, so they remain largely unfamiliar to potential users.

This has caused many portals to become mere repositories of data rather than potential wells for data diving. Despite sources are growing rapidly, the extensive exploitation of them is still in its childhood.

As pointed out by Gascó-Hernández et al. (Gascó-Hernández, Martin, Reggi, Pyo, & Luna-Reyes, 2018) this is also largely due to a lack of technical skills and training to exploit the value of open data.

However, this lack of skills could be in turn related to the many issues presented by these resources (Schintler, 2014). Beyond the volume and the complexity of the available data, the velocity and veracity (uncertainty) of the data sources affect their quality, accuracy and completeness (Miller & St, 2013) (Hemerly, 2013).

This extensive availability of data[11] with vast unused potential and informative power is the starting point of this document that seeks to exploit the high value of this information.

The four main sources of open data used in this research are[12]:

- ESIF Open Data Portal;
- Smart Specialisation documents;
- RegPat OECD database;
- CORDIS Research and Innovation database.

[11] Other open data platform and REST API:
http://data.europa.eu/euodp/en/developerscorner
https://www.europeandataportal.eu/en/
https://www.dati.gov.it/content/sviluppatori
https://bdap-opendata.mef.gov.it/
https://opencoesione.gov.it/en/api-opencoesione/
http://www.agenziacoesione.gov.it/it/arint/OpenAreeInterne/index.html
Agenzia Italia Digitale: https://developers.italia.it/it/api

[12] Also Eurostat data via the web service endpoints have been merged to the core set of data as explained in chapter 5 (http://ec.europa.eu/eurostat/web/json-and-unicode-web-services)

This document is organized in two main parts:

- Firstly, a comprehensive overview of the use of open data at European level for monitoring and performance assessment. Using ESIF open data, the research focuses on developing a wider and deeper approach to the use of open data for simpler and more effective interpretation and insights;
- The second part presents additional sources of open data (RegPat patents and CORDIS projects) in the perspective of strategy development and assessment. The research topic of this second part highlights some ways open data can be used for data-driven policy making as regards the innovation strategy for R&D investments.

2 Open data for performance monitoring and assessment

2.1 Technologies and monitoring tools at EU level

The EU system for monitoring EU programmes implementing ESIF funds is based on different tools and technologies.

The system enables monitoring of ESIF programming and implementation to support the strong result-orientation approach of the 2014-2020 European legislative framework.

It is mainly based on two web applications; the System for Fund data management of the European Union (SFC) and the ESIF open data portal.

The first application is a management tool for ESIF managing authorities (MAs) at national and regional level and the ESIF managing Directorates General: REGIO, EMPL, AGRI, MARE, HOME.

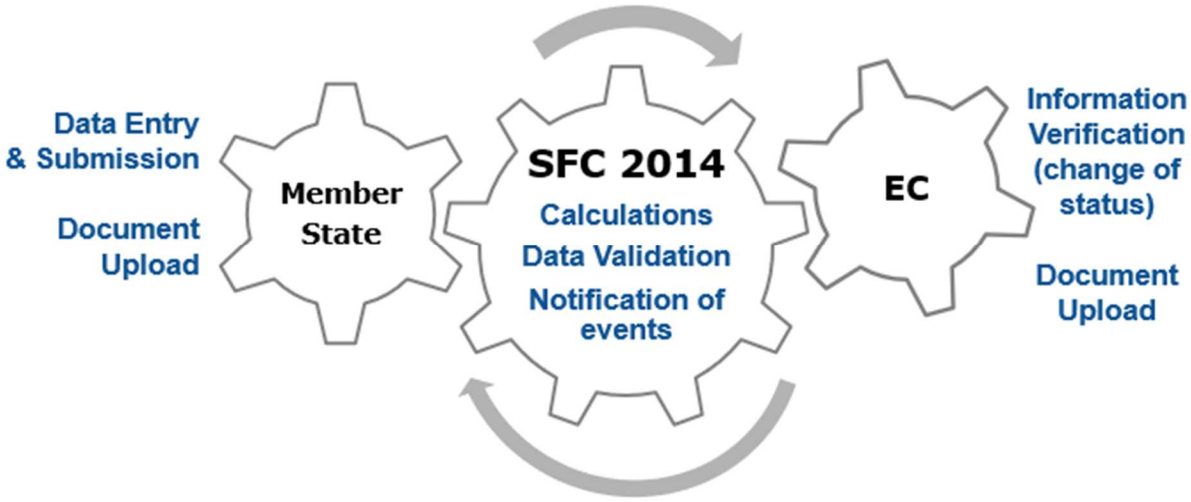
SFC's main function is the electronic exchange of information concerning shared Fund management between Member States and the European Commission as described in Article 74(4) of Regulation (EU) No 1303/2013.

In other words, the tool is mainly dedicated to programme data input by MAs of programmes implementation, verified by Directorates Generals and, if needed, corrected by the MAs.

National and regional MAs of financial programmes have specific deadlines to report financial and output information using a data structure detailed in the Current Provision Regulation and in fund specific regulations.

A view of the SFC portal is in the following figure.

Figure 6 – The EU system for fund data management (SFC)



Source: https://ec.europa.eu/sfc/en

Most of this information is then available to the general public for consultation using a web platform dedicated to data consultation and data visualization as presented below.

2.2 Monitoring ESIF: regulation and information reporting

The ESIF open data portal gives access to EU Cohesion Policy data, one of the western world's largest collection of local and regional development policies operating under a single legal and institutional framework. It targets all EU regions and cities and is aimed at fostering competitiveness, economic growth and new jobs in regions.

The planned resources over the 2014-2020 period for different funds are almost EUR 650 billion with EUR 460 billion of EU resources. The use of resources is regulated by the Common Provision Regulation[13] (CPR).

Each European region can access planned financing in the different funds according to specific investment strategies defined within their Operational Programmes (OPs).

These documents have a predefined structure for selecting and categorizing investment decisions and for reporting each year. In particular, data communication to the European Commission is regulated in the CPR under Article 112, regarding the transmission of financial data and Article 72(d) as regards the systems for accounting, storage and transmission of financial and indicators data for monitoring and reporting.

Selected projects, declared expenditure, output and result indicators must be monitored by the MAs responsible for the management of funds and reported to the centralized information system within the Annual Implementation Report (AIR) for European Commission approval.

ESIF data in the Open Data Portal covers more than 540 OPs under the five ESI Funds[14]. Data are available in many financial datasets which are related to planned, implemented and paid resources. Data on selected common indicators, targets and implementation relative to the actual output of deployed financial resources are stored in the achievement dataset.

Intended users of this data include anyone interested in monitoring policy development, especially EU citizens, Member State administrations, EU Institutions, policy makers, researchers and practitioners in regional development studies.

[13] Common Provision Regulation (EU) No 1303/2013 of the European Parliament and of the European Council of 17 December 2013 laying down common provisions on the European Regional Development Fund, the European Social Fund, the Cohesion Fund, the European Agricultural Fund for Rural Development and the European Maritime and Fisheries Fund and laying down general provisions on the European Regional Development Fund, the European Social Fund, the Cohesion Fund and the European Maritime and Fisheries Fund and repealing Council Regulation (EC) No 1083/2006 https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32013R1303&from=EN

[14] European Agricultural Fund for Rural Development (EAFRD), the European Regional Development Fund (ERDF), the European Social Fund (ESF) with distinct data for the Youth Employment Initiative, the Cohesion Fund (CF) and the European Maritime and Fisheries fund (EMFF).

There are an increasing number of EU open data discussions and initiatives at local, regional, national and EU levels with public events, workshops, conferences, presentations and use of data events such as hackathons and datathons.

These initiatives have multiple objectives and potential benefits including increased transparency and accountability, more efficient communication to citizens and journalists, supporting debates on policy performance and fueling a new data-driven decision mindset for policy makers.

For the **performance** of regions in implementing their OPs the two most important financial measures are project selection (resources allocated to investments) and expenditure declared (resources disbursed to beneficiaries) as reported by the MAs.

The progress and performance of each OP is monitored against the financial amount decided during the planning phase which was at the beginning of the programming period in 2014.

Data are available disaggregated by regulation categories:

- Fund
- OP
- Priority Axis
- Thematic Objectives (i.e. the macro priorities for investment)
- Fields of intervention (i.e. the micro priorities for investment)
- Category of regions (more developed, less developed, transition).

While the planned financial amount can only be updated within a reallocation of OP resources, the financial implementation data are updated three times per year, at the end of January, July and September.

Implementation data submission for common indicators is scheduled at the end of each year. Indicator targets (planned) are not subject to variations unless there are OP modifications.

According to the API paradigm, the platform exposes each of the above datasets using dedicated web services. The information in each dataset is organized in a clear JavaScript Object Notation (JSON)[15] structure where each regulation dimension is the JSON *key* and the OP data is the specific value. The endpoint is referenced by a unique code and is accessible using an HTTP request from a browser[16] or from the most common data analytics programming languages (e.g. R, Python, Javascript, etc.).

[15]JSON is a lightweight data-interchange format http://www.json.org/

[16] For an example see https://cohesiondata.ec.europa.eu/resource/f6wa-fhmb.json

The very large amount of information, complexity of data structure and frequent updates mean that advanced and agile tools are needed to easily fetch, parse, analyse and visualize information instantaneously.

However, many users primarily interested in policy progress find it difficult to explore and easily extract, transform and analyse the data. Visualizing up-to-date data in the form of charts, tables and other infographics, simplifies exploration of programme performance. This enables researchers, policy makers and the general public to assess the effectiveness of programmes in effectively deploying taxpayer money without data diving and regardless of their skillset and expertise level.

These aspects should be carefully considered, as noted by Gascó-Hernández et al. (Gascó-Hernández et al., 2018). Despite the potential transformative value of open data when they are made more discoverable, accessible and available in alternative formats, there is limited evidence of actual use. This is partly attributable to the lack of fundamental expertise and technical knowledge related to data management and visualization (Graves; Hendler, 2014). Most importantly, many users are not even aware of the data potential, its possible use and the technological and analytical impact (Ramon Gil-Garcia, 2017).

3 Monitoring EU programmes implementation using data visualization

3.1 State of the art for visualizing monitoring data

The increasing and heterogeneous group of data users deciding and debating programme implementation of structural funds require information dissemination based on generally understandable concepts.

For this, visualizations are the easiest and fastest tools for human eyes to see and recognize patterns and trends.

However, the huge amount of multi-dimensional information raises the problem of successfully and easily stimulating visual reasoning using relatively simple tools to synthesize data. Researchers are adopting new tools and technologies to analyse increasingly large economic data sets generated in greater volumes.

Big data often offers valuable information to be extracted and interpreted but the time when simple bar charts or scatter plots were enough is long gone. Thus, the development of advanced data visualization techniques is becoming a necessary and challenging area of research and interest.

Data visualization can help in making sense of large data sets by presenting contents in an innovative visual format that does not require multiple tables, or lots of rows and columns. Furthermore, the connection between several data sources generates newer and larger datasets leading to further discovery and information.

However, there is an increase in the complexity and volume of data that is collected, stored and made available by institutions and public bodies. Literature shows that open data government datasets still have several barriers including inadequate collection, classification, processing and presentation tools, non-standardized data description and formats, as well as missing or incoherent data. This makes it hard for different users and analytical approaches (Dawes & Helbig, 2010).

IT investments and skills devoted mainly to storing systems, architecture, software, hardware, security, networks and Web technologies without an explicit purpose for data exploitation are poorly suited to the new paradigm for using data as an asset for BI and data science. This in turn affects the benefits of open data initiatives and sharing, especially at even lower levels such as local administrations and municipalities.

As a response to the challenges of managing vast amounts of government data and making it accessible for different purposes and informational needs, Dawes explains the concepts of stewardship and usefulness. Among the 'stewardship proposals' to improve he suggests creating and improving metadata for each data source, improving the data management system and adopting standard data formats.

As 'usefulness proposals', he suggests providing easy-to-use basic features as well as improving and enhancing searches and displays of data (Dawes, 2010). Noveck (Noveck, 2012) adds that it is also important to have high quality standards for dissemination catering to different needs and uses by citizens and other social actors. Merino et al. (Merino Huerta, Mauricio, 2010) consider the delivery of public data as opportune and reliable for better decision making in government as well as for government accountability concerning public decisions and actions.

The use of different technology tools to implement open data initiatives is recognized as a 'fit-the-right-tool-for-the-job' situation. So each complex economic, social and political issue along with the data it generates relates to different approaches and methods for information production and use (Birkland, 2014) (McCool, 1995).

Government open data across different end-users is available with intense use of technology such as IT tools and Web applications (Dawes & Helbig, 2010) (Noveck, 2012). IT tools and web applications are currently the engine of the debate concerning open data as they can both provide the 'raw material' for different types of users as well as receive new information and data from those users; decision makers, analysts, researchers or citizens (C.Hood, 2007).

There is a wide range of technological tools available for policy analysis and data visualization. Flexible and powerful information technologies and various analytical methods are supported by several open data initiatives. This scenario is constantly evolving, but a brief overview of current tools and platforms used to visualize and analyze open data is below.

Open Data Portal for ESIF

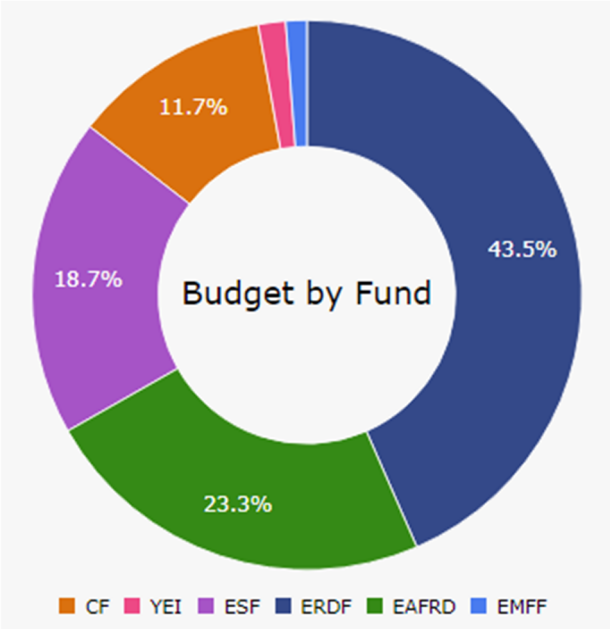
The Open Data Portal for ESIF[17] proposes a visualization tool for broad aggregations of data at Member State, Fund and Thematic Objective levels in terms of planned, implemented and paid amounts.

Compared to the 2007-2013 programming period when information was only shared in spreadsheets and documents, the tool significantly improved accessibility and reuse of data for accountability and transparency. The tool allows the user to easily shift among the menu views to explore real time data on programme implementation and budget.

In the following figures are examples of the proposed data visualization and the budget by fund shows the aggregation of planned financial resources by fund highlighting the large share of ERDF within the policy.

[17] https://cohesiondata.ec.europa.eu/overview

Figure 7 – Budget by fund, percentage of total

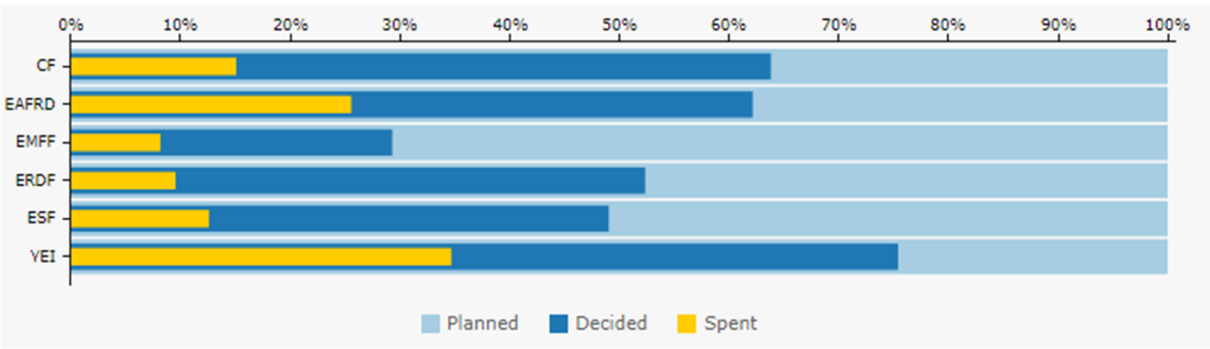


A different view (Figure 8), based on the fund dimension and financial implementation makes it easy to understand the progress by fund in terms of resources allocated and disbursed.

Currently, the Youth Employment Initiative has the highest share of both resources allocated (decided) and disbursed (spent).

However, when combining this information with the previous figure, Youth Employment Initiative planned resources are only a limited amount of the full budget, so they are more easily allocated.

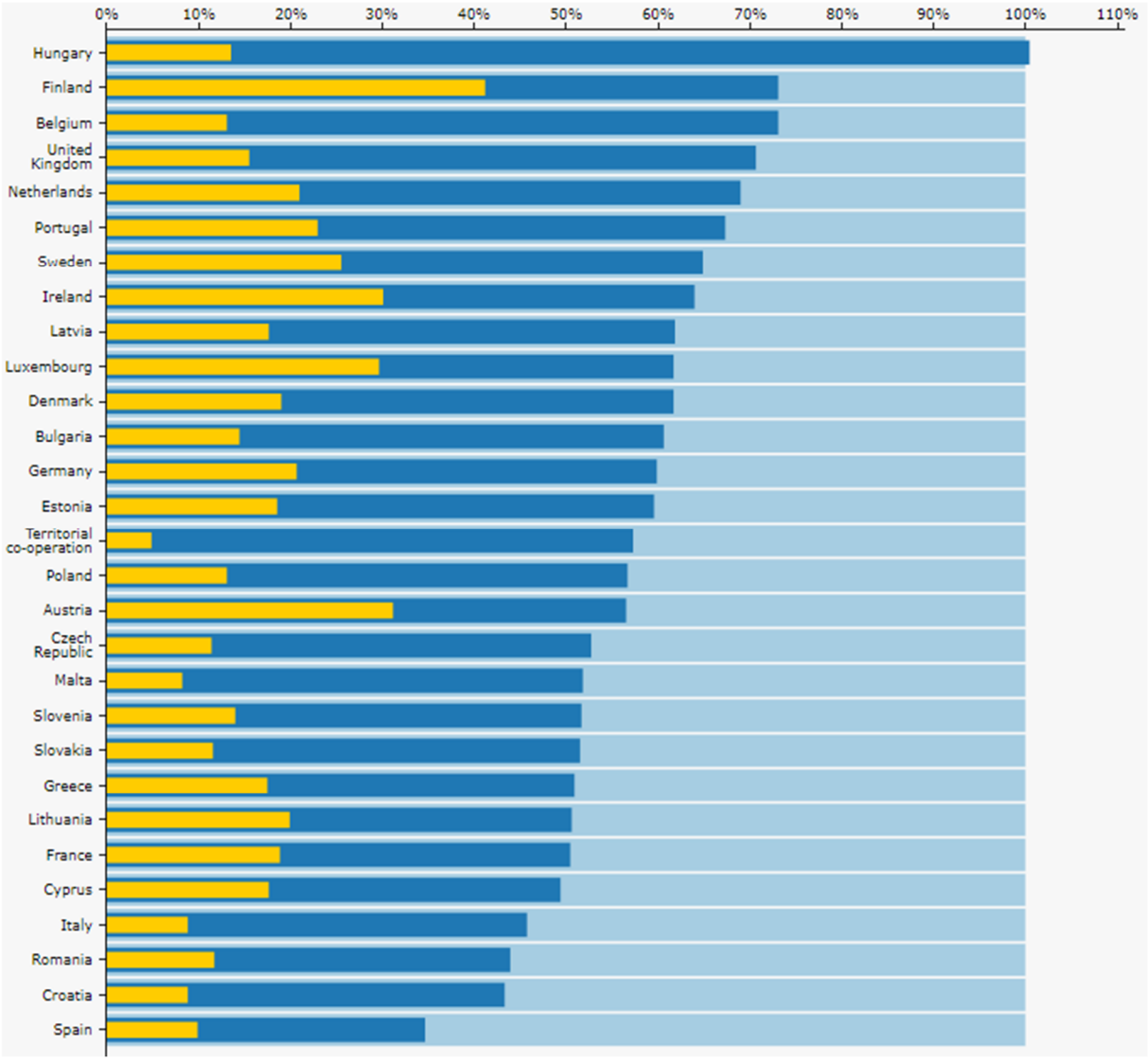
Figure 8 – Implementation progress by fund, share of planned



This suggests the importance of a multi-dimensional approach combining more variables in the same analysis dashboard.

A similar graphical solution adopts the Member State as the aggregation dimension and orders by size to show a ranking.

Figure 9 – Implementation progress by country, share of planned



Many other visualizations on the platform cover the main variables of interest, providing a clear overview of the status of implementation and highlighting regional authority efforts. However, much of the data potential remains unexploited as only macro dimensions (e.g. Fund, Member State, Thematic Objective, years) are used for data aggregation and comparison.

This fundamental issue has driven the research and development of higher detail visualization supported by enhanced visualizations as presented in chapter 4.

Furthermore, despite the accessibility of data (in many formats) customized visualization tools (e.g. Plotly) are unlikely to be used due to the technical skills and domain specific knowledge needed (Gascó-Hernández et al., 2018).

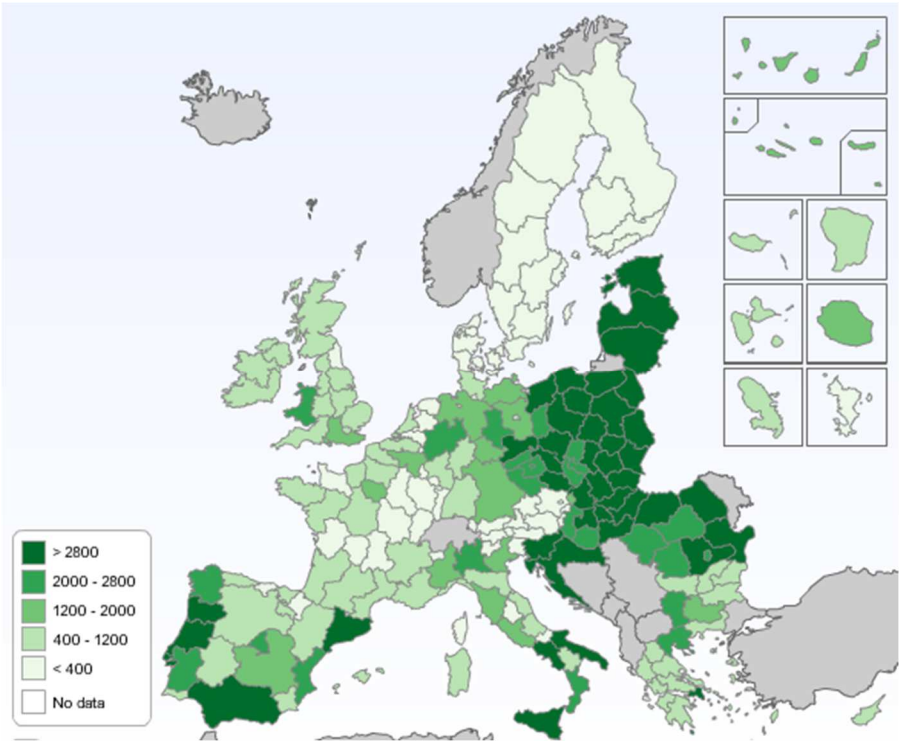
ESIF viewer

ESIF - Viewer[18] is a tool to search planned investments in ESIF data (ERDF, CF, ESF and Youth Employment Initiative) and contains data from ESIF OPs. The amounts are presented at regional level and include data from regional OPs, but also shares of national and transnational cooperation programmes. The user can search for planned investments per country, region, OP-type and different categories of intervention.

For a performance assessment approach, this tool has drawbacks:

- No real time data: currently the tool contains data from ESIF OPs retrieved on 20/01/2017 from the SFC2014 database;
- Planned data: only planned financial resources are presented in the current plot and table, without reference to progress and achievement indicators;
- Estimated data: the proposed visualization includes data from regional OPs, but also shares of national and transnational cooperation programmes. The total shares have been estimated by taking into account the population size of the regions. Therefore, these are estimates and do not reflect precise investment figures.
- Unique visualization: despite the powerful geo representation, data are represented using only one map for each selection.

Figure 10 – Regional planned investments in ESIF-viewer



[18] http://s3platform.jrc.ec.europa.eu/esif-viewer

ICT Monitoring

While the ESIF viewer covers planned investments overall, ICT Monitoring[19] contains data from ESIF OPs on planned ICT related investments. The amounts in this tool are presented at regional level using a unique map. Users can search in three dimensions (amounts, keywords and financial forms) and four categories (Member State, Region, Thematic Objective and Categories of intervention).

As with the ESIF viewer this tool has updating and estimation drawbacks as well as a similar approach to presenting planned data in a unique view.

R&I Regional Viewer

R&I Regional Viewer[20] enables visualization and comparison of planned Research & Innovation investments under different funding channels (i.e. ESIF and Horizon 2020) across EU Regions. The tool also combines financial data with Eurostat data sources, showing regional economic indicators (GDP, population, R&D, unemployment) as well.

Although the tool enlarges and combines the set of data sources, its main topic and purpose mean that ESIF indicators are limited to resources dedicated to R&I. As with similar tools, Regional Viewer shows estimated planned data with a relatively large lag in data updates.

Monitoring Helpdesk project

A BI tool based on data visualization is helping EC geographic units and desk officers to monitor advancement of the programmes. The tool implements a deeper level of financial and indicator analysis with a wide set of data visualizations to assess each programme progress weekly with updated information.

The tool uses additional categories such as the form of finance (e.g. grant, loan, equity, etc.) and the intervention fields (e.g. research and innovation infrastructure, technology transfer, SME business support, advanced support service).

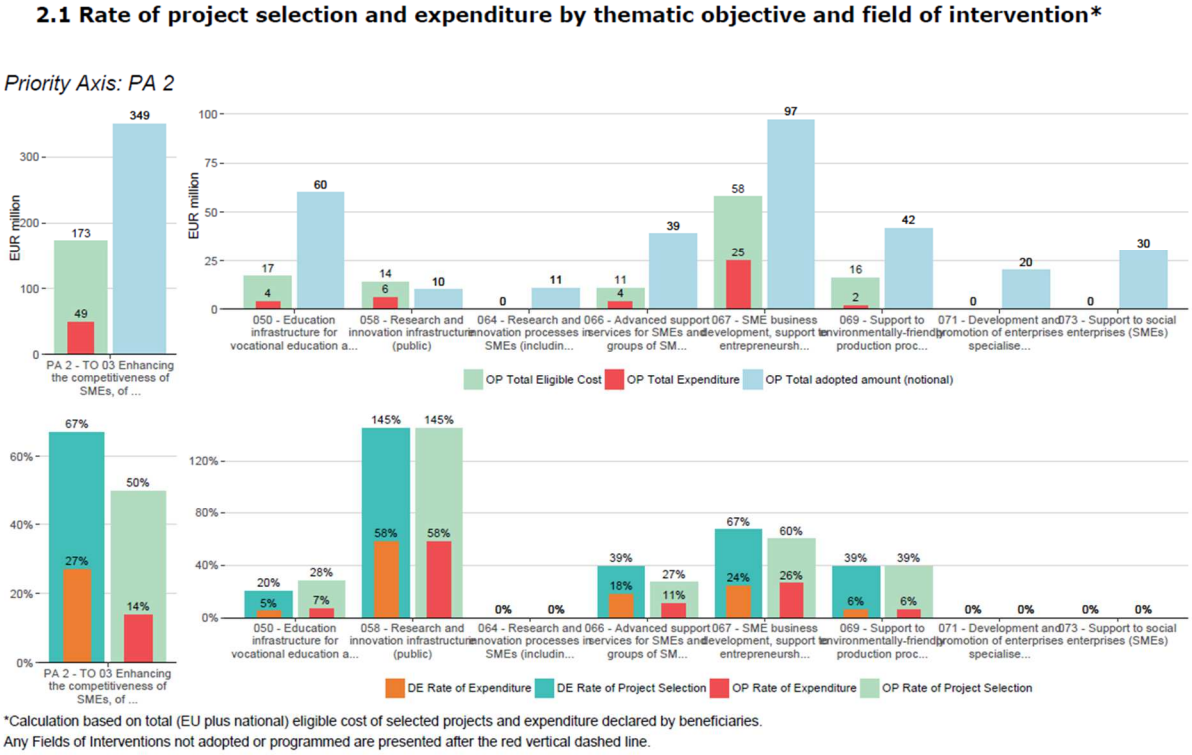
Furthermore, it covers additional variables and indicators including certified expenditure, resources paid and physical indicators.

Attention is also paid to comparison between years and programming periods as well as to the forecast and expected performance.

The following figure shows an example of the view using a combination of many variables aggregating planned, allocated and spent financial resources.

[19] http://s3platform.jrc.ec.europa.eu/ict-monitoring
[20] http://s3platform.jrc.ec.europa.eu/synergies-tool

Figure 11 – OP monitoring



Source: own elaboration

OpenCoesione

Despite the significant performance reporting and assessment tools at EU level, current regulation covers only programme level with no requirements for submission of further details at the deeper **project level**. Micro data at MA level on financed projects, aggregated for AIR submission is undoubtedly the most fundamental asset for monitoring and decision making as it gives access to crucial information at ground level.

Data on the types of beneficiaries, economic sectors, average size of projects, duration of implementation and geographic localization could exponentially increase the informative capacity of data and advise policy.

Furthermore, these data lead to an easy expansion of the information base. For example, including geolocalization and sectors enables combinations with information from other sources using these variables as sort of foreign key for joining.

Currently, the Italian OpenCoesione is the only platform in the EU sharing information on Italian projects financed through cohesion policy resources. It covers almost EUR 100 billion of funding and almost 1 million projects over the 2014-2020 period.

This web application is organized as a powerful BI dashboard with a menu for data exploration that triggers several charts and visualizations that can be updated.

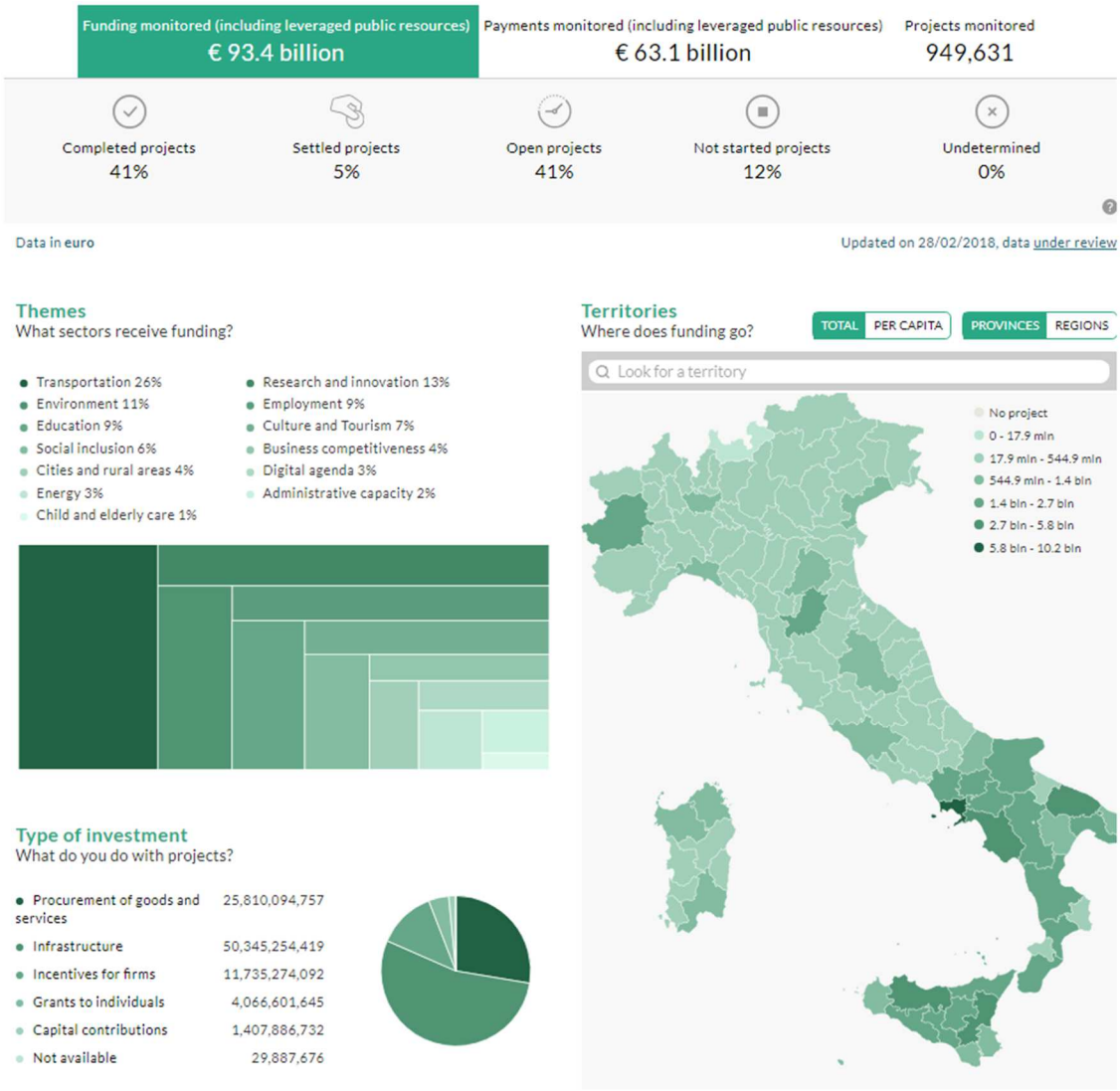
The main view shows an overview of the variables under three dimensions:

- Themes: broad sectors (e.g. transport, environment, R&D);
- Territories: Italian regions (NUTS2) and counties (NUTS3);
- Type of investment: project action (e.g. infrastructure, firms support, etc.).

Furthermore, in line with advanced BI tools, filters, drill-down and roll-up operations are enabled with these three dimensions to which can be added the fund and beneficiary.

For accountability and transparency, data are also available in common exchange format and exposed via API[21] for reuse in applications. The OpenCoesione API is an application interface that allows any external software component to access OpenCoesione data on projects and entities financed by cohesion policy in Italy.

Figure 12 – Project level monitoring: OpenCoesione (IT)



Source: https://opencoesione.gov.it/it/

[21] https://opencoesione.gov.it/en/api-opencoesione/

Beyond the specific domain of ESIF financial resources and programme monitoring, there are other tools at EU level that consider broader economic and financial aspects.

OpenBudgets

OpenBudgets[22] is a Horizon 2020 project focusing explicitly on corruption and the comparison of budgets between administrative regions and other government levels.

This web platform offers a toolbox to everyone who wants to upload, visualize and analyze public budget and spending data.

It has easy to use visualizations and high-level analytics along with fun games, accessible explanations of public budgeting and corruption practices as well as participatory budgeting tools. It caters to the needs of journalists, researchers, policy makers and citizens.

Regional Benchmarking

Regional Benchmarking[23] is an interactive tool for Regional Benchmarking which helps identify structurally similar regions across Europe through statistical indicators that cover social, economic, technological, institutional and geographical characteristics. The objective of the tool is to identify regions with similar characteristics to foster cross regional cooperation and the exchange of knowledge, especially on innovation.

EU Trade

EU Trade[24] is a fully interactive web-based application to visualize and analyze inter-regional trade flows and the competitive position of regions in Europe. This tool makes it possible to assess regional assets and to analyze a region's economic position. This is a first, fundamental step in the process of building place-based and evidence-based regional policies and smart specialisation strategies.

Other Horizon 2020 projects

The Crowd4Roads[25] project aims at engaging drivers and passengers in the development and adoption of more sustainable car usage habits and road maintenance policies. It is based on SmartRoadSense, a crowd sensing system which uses the accelerometers of car-mounted smartphones as non-intrusive sensors of road surface quality. This generates open data for an aggregated road quality measure shown in a geographical map[26].

[22] https://openbudgets.eu/
[23] http://s3platform.jrc.ec.europa.eu/regional-benchmarking
[24] http://s3platform.jrc.ec.europa.eu/s3-trade-tool
[25] http://www.c4rs.eu/
[26] http://www.smartroadsense.it/data/map/

Your Data Stories[27] is a platform that helps make sense of open and social data. It looks to better satisfy the needs of the 'demand side' – meaning citizens, journalists and others with a better 'supply' of open data (traditional producers and user-generated content). YDS addresses professionals in government, public administration, business and journalism, but is also made for citizens.

ROUTE-TO-PA[28] is a multidisciplinary innovation project that combines expertise and research in e-government, computer science, learning science and the economy. It aims at improving the impact, for citizens and within society, of ICT-based technology platforms for transparency.

Digiwhist[29] looks to increase trust in government, improving the efficiency of public spending across Europe by sharing information. The systematic collection, structuring, analysis and broad dissemination of information on public procurement and mechanisms aims at increasing the accountability of public officials across the EU and in some neighboring countries.

Smarticipate[30] gives citizens access to data about their city in an easy to understand way, enabling them to better support the decision-making process. Local governments will be able to tap into the ingenuity of their residents, gaining valuable ideas. This two-way feedback makes cities more democratic and dynamic. Residents will also play an active role in verifying and contributing to data.

3.2 Data visualization theory

The natural behavior of human beings before taking decisions is to acquire information.

In an information technology process, this short sentence would delineate a situation in which there are two types of interacting classes of objects and where is of crucial importance the definition and quality of their attributes. On one side, enough level of cognitive capacity and skills is required on the human side, and correctness and timeliness of the information stands on the other side.

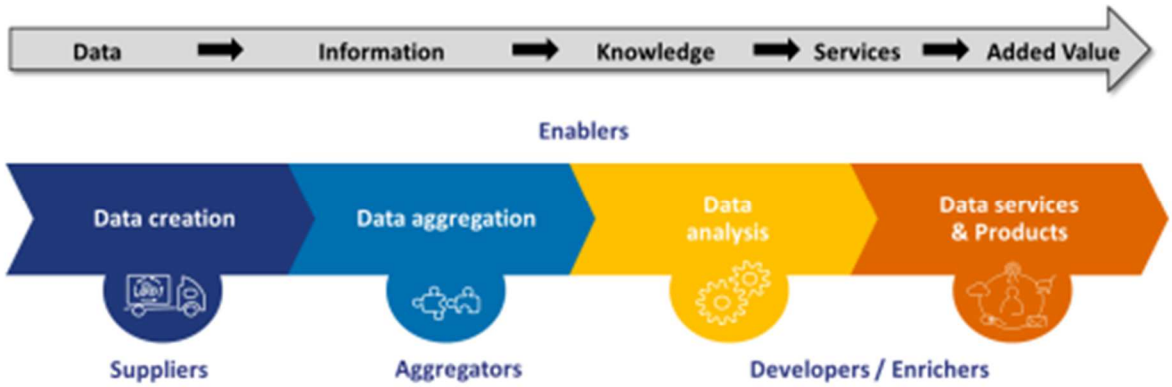
However, the two attributes follow different pattern of growth with small and slow acquisition of skills for human beings and an exponentially large and fast amount information available. Thus, despite correctness and timeliness of information, a new attribute of information is increasingly becoming important in modern decision science.

The main reason behind the transformation of data into graphical images is that is far more **time-saving** to get knowledge from depictions than looking through text and numbers.

[27] https://yourdatastories.eu/
[28] http://routetopa.eu/
[29] http://digiwhist.eu/
[30] https://www.smarticipate.eu/

Millions of lines of raw data even if presented in a table would not tell nothing about the information they hide unless even the simplest figure aggregates and plots them all.
Since it saves time, it could be encompassed within the realm of data services, final output of the data value chain (Berends, Carrara, Engbers, & Vollers, 2017).

Figure 13 – The data value chain



Source: Re-using Open Data (Berends et al., 2017)

As the information generated in the last years has been continuously soaring, decision makers have to access an increasing information amount per time unit. This could happen only synthetizing information in graphical form, as opposed to a tabular or textual form.
Hence, data visualization encompasses all the set of techniques and tools for the acquisition, processing, transformation and communication of raw data into useful knowledge to satisfy an information need.
Data visualization is a quite new and promising field in computer science flourishing due to the growing data available in any field and, in parallel, to the changing mindset towards a more data-driven decision-making approach.
This changing approach relying on data for better decision-making supported by improved technologies and advanced systems to carefully craft messages is underpinned by the concept of **data culture** (Giest, 2017).
It represents the widest used BI tool for discussion and decision-making as it is usable and understandable by heterogeneous audience both in terms of size and skillset. It helps engaging more diverse audiences in the critical process of analytic thinking of quantitative and qualitative variables.
Modern techniques and algorithms for creating effective visualizations are based on principles from graphic design, visual art, perceptual psychology and cognitive science.
Computer science plays the rule of process enablers using computer graphic effects to reveal the patterns, trends, relationships out of datasets.
Graphical representation allows decision makers to see analytics visually, to quickly grasp the concepts and insights relevant for the development of strategies and corrective

measures. It encourages appropriate interpretation, selection and association stimulating human senses and cognitive processes for pattern recognition, comparison and analysis.
For example, the analysis and comparison of financial data could be an extremely complex task if not accompanied with an effective graphical representation able to synthetize the stream of data produced, especially when in real-time.
A fundamental question in the visualization is what constitutes an effective visualization within the domain of analysis. Despite some concepts and rules apply in general in the development of charts and graphical representation, the theory of data visualization drives the process of effective visualization production through some specific concepts.
According to Ward et al. (Ward, Grinstein, & Keim, 2010) *[…] to create the most effective visualization for a particular application, it is critical to consider the semantics of the data and the context of the typical user. By selecting data-to-graphics mappings that cater to the user's domain-specific mental model, the interpretation of the resulting image will be greatly facilitated. In addition, the more consistent the designer is in predicting the user's expectations, the less chance there will be for misinterpretation*.
First, a fundamental aspect of data visualization is the knowledge of the **audience** and the understanding of how it processes visual information, even referring to tools already adopted and shared. At the same level of importance is the knowledge of the specific **domain** and the variables to be visualized.
Thus, the second point refers to the process of data analysis oriented towards specific questions and issues of investigation for the specific audience. Only the true and deep understanding of the variables under analysis and the nature of the information of interest (i.e. the domain) can drive the effective representation of data.
In this sense, data visualization should be considered a topic-driven and audience-driven formatting process.
As the objective of the visualization is to give those caring about the topic the greatest number of concepts and information in the shortest time within a finite space (e.g. pdf page or a web page), the selection of the most important results to be displayed coming out from the data analysis should be based on:

- Focusing the attention and alerting on specific facts;
- increasing the understanding and awareness of the fact;
- simplifying the remembering of its main points.

Thus, the representation should immediately reveal patterns and peculiarities of data (e.g. trend, relationship, outliers, errors), organize complex information in a way accessible/tailored on the pertinent audience and highlight concepts to immediately remind in the aftermath of the discussion/presentation.

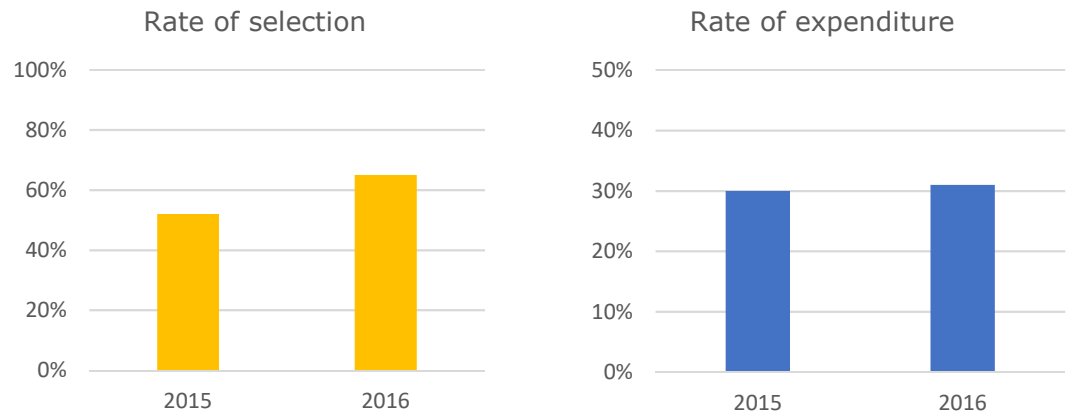
Third, the presentation of visualization should follow an order, usually form the general to specific, and reveal the data gradually in order to avoid confusing and overloading the viewer. This avoid the viewer an excessive cognitive processing task, keeping in mind the root of the issue under discussion while drilling down toward additional information and finally to the conclusion of interest. Conveying a narrative with visualizations often requires choosing an order in which to present visualizations. While evidence exists that narrative sequencing in traditional stories can affect comprehension and memory, little is known about how sequencing choices affect narrative visualization (Hullman et al., 2013).

Furthermore, visualization **misuse** and the harmful effect of driving audience towards misrepresentation, disinformation and even deception has to be carefully considered (Cairo, 2015) (Pickle & Monmonier, 1997). It should be noted that here the discussion refers to the visualization bias based on correct data, different from statistically biased underlying data as largely discussed by Schintler et al. (Schintler, 2014).

Aspects of substantial importance related to graphical perception have to be taken into account when designing outputs as for example:

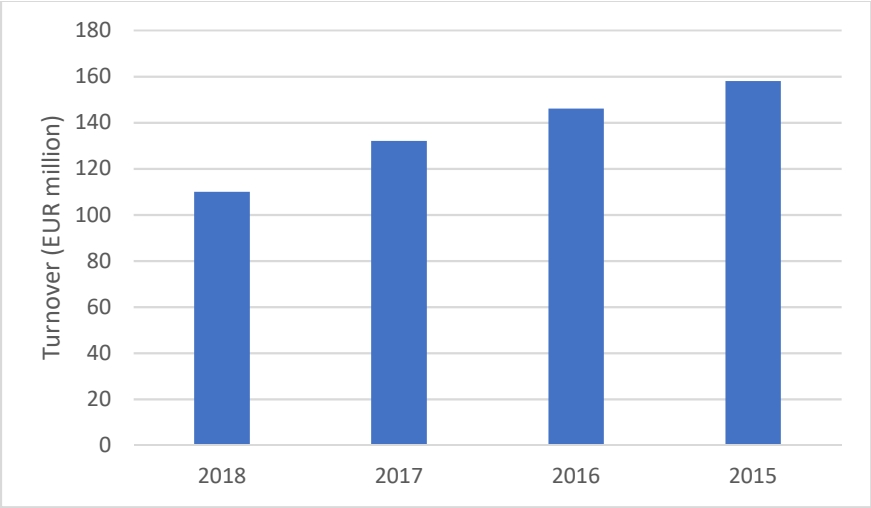
- scale bias: perspective distortion (Tufte, 2001) using the same proportions for comparable variables. The following figure creates the false impression of similar percentage level, but the rate of selection is relatively more advanced than expenditure. The absence of columns label on top does not help the reader as well.

Figure 14 – An example of scale bias



- meaningful order: the order of the visualization elements could improve the perception and make the message immediate whereas mixed elements complicate the interpretation. The following figure induces a misleading message (positive trend) because of the order;

Figure 15 – Time moves forward

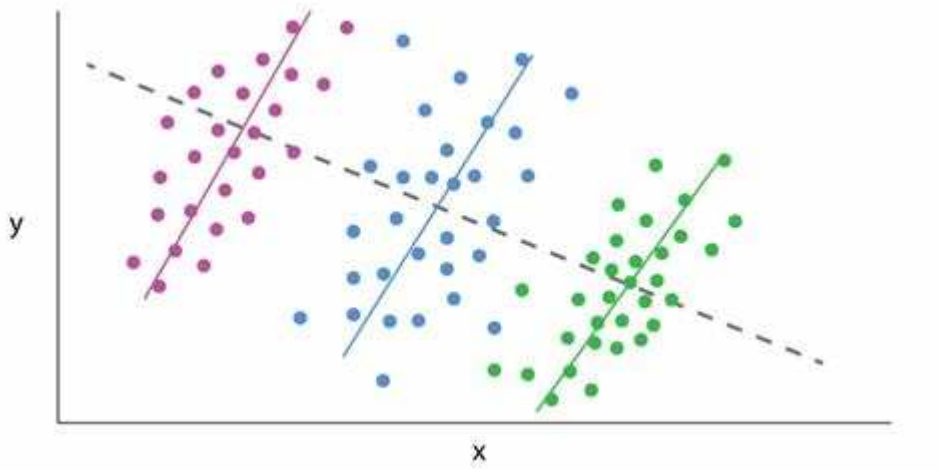


- decoration consistency: if the visualization presents several graphics, colors must be consistent, e.g. the same colors represent the same aggregation across the different views;

- aesthetic moderation: the excessive use of grid, labels, notes, colors could lead to what Tufte (Tufte, 2001) indicated as *chart junk*, supplementary components useless to the interpretation;

- Simpson's paradox: aggregation applied produces a certain trend that is reversed when a different aggregation is used (e.g. different subgroups);

Figure 16 – A Simpson's paradox example



- Stacking bias: confusion in the perception of the element origin of values as it is unclear if the origin is in the axis level or stacked above another element.

Hullman et al. (Hullman, Adar, & Shah, 2011) found that biased signals lead to biased interpretations with consequent unproperly decision process.

On the opposite, according to data visualization literature (Bateman et al., 2010) (Borkin et al., 2013), charts aesthetic factors may play a major role as regards comprehension and memorability of the message.

This poses several research questions on the development of the output **concept**:

- ✓ how the visual factor should be appropriately chosen given the topic under analysis?
- ✓ What is the best visual solution to properly convey the message of interest?
- ✓ How to synthetize the largest set of data available?

The following figures are an example of improved data visualization in terms of function and form, developed according to the theoretical principles examined and discussed in chapter 4.

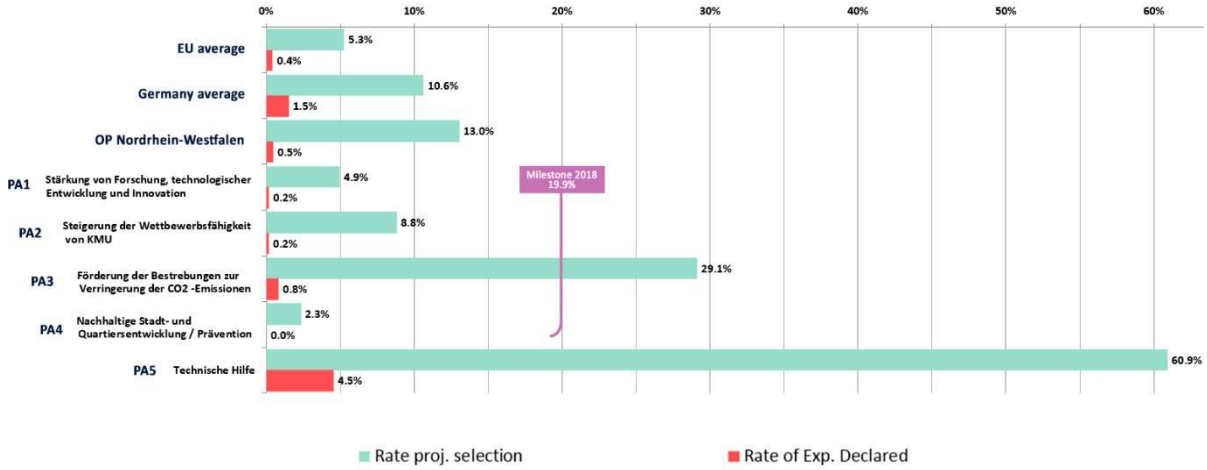
Both these figures compare the rate of selection and expenditure by Priority Axis of the OP with reference to the overall OP level, the MS and EU progress. They only differ substantially in the use of aesthetic factors from (a) to (b).

The main elements of improvement between the two, functional to the analysis and interpretation of data, are:

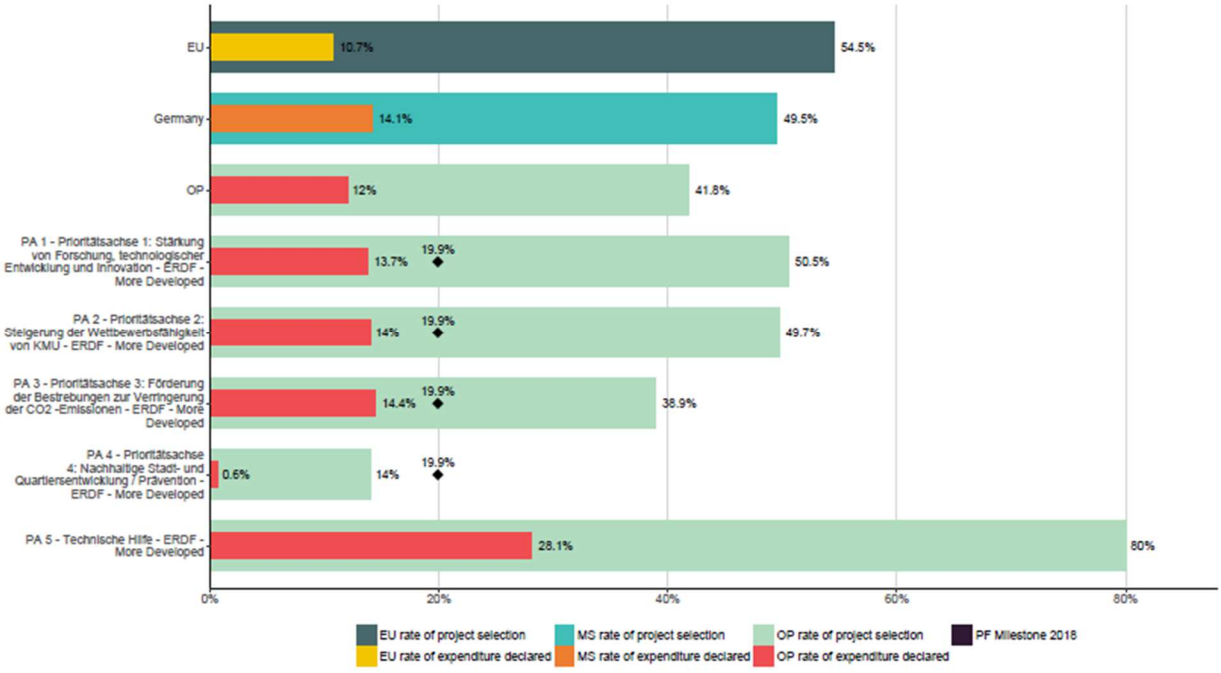
- ✓ The adoption of nested variables to emphasize the selection and expenditure relationship given by the specific application domain. An important aspect is to avoid comparison only among the figure bars of the same colours (i.e. red with red, green with green) but also among red and green bars;
- ✓ Directly related to the previous point is the possibility to indirectly bind business rules into visualization. The relationship between selection and expenditure implies that green bars shall always be greater than red allowing to easily spot any data errors;
- ✓ The adoption of colours to reflect the different geographical levels represented as not immediately perceived with a unique colour scale. In the view of a benchmarking framework, the use of colours allows to better emphasize the comparison based on the spatial dimension;
- ✓ Minimize the size of accessory but not central variables (milestone). Despite all the variables proposed in the image are important, they are not important the same way. This implies that some variables with a marginal role should have a different representation and importance in the output. In the example, despite the milestone conveys the same message, its relevance within the figure is reduced.

Figure 17 – Substance and form: improving interpretation

(a)



(b)



In the design and prototyping phase, these elements have to be considered when defining the aspects of interest to be conveyed through the image. Further details are presented in chapter 4.

3.3 Interactive visualizations

The informative power of visualizations conceived according to the theoretical principles, audience and domain could be further improved and enlarged adopting dynamic representations (Ward et al., 2010).

There are several reasons why interactive data visualizations are better than static:

- **More information**. Interactivity allows users to embed much more information than in a static visualization by using selection, filters, tooltips, click-events on dimensions and measures;
- **Easier perception**. The possibility to show some dimensions and measures only on demand and hide when not needed will allow users to focus only on the details they are looking for.
- **Gamification.** The possibility to actively browse information of interest without being a mere passive spectator of static views. Responsive data encourages users to explore more and, subsequently, receive more insights.

According to Ward et al. (Ward et al., 2010) interaction within the data and information visualization context is a mechanism for modifying what the users see and how they see it. As described by Yi et al. (Yi, Kang, Stasko, & Jacko, 2007), a broad classification of existing interaction techniques could be as follows:

- Selection - user controls for identifying an object, a collection of objects, or regions of interest to be the subject of some operation, such as highlighting, deleting, and modifying. Decisions need to be made on what the results should be for a sequence of selections and its level of granularity;
- Filtering - user controls for reducing the size of the data being mapped to the screen, applying constraints to records, dimensions, measures.
- Navigation (exploration) - used to search for a subset of data to be viewed, the orientation of this view, and the level of detail (zooming).
- Reconfiguring - user controls for changing the way data is mapped to graphical entities or attributes, such as changing the scale dimensions, reordering the data or layouts, or transform the data.
- Encoding (aesthetics) - user controls for changing the graphical attributes, such as point size or line color, to potentially reveal different features.
- Connecting (split up views) - user controls for visualize many different linked views or objects to explore possible related items.
- abstracting/elaborating—user controls for modifying the level of detail.
- hybrid—user controls combining several of the above in one technique.

An effective dynamic visualization largely depends on the capacity to anticipate the types of views and view modifications that will be of most use to the typical user of data, and then provide intuitive controls for setting and customizing the views accordingly.
Each supported view should be intuitive from the set of controls available and selecting a new view should require minimal actions on the user's part. As useful views depend heavily on the type of data being presented and the task associated with the visualization, the clear understanding of the domain and data is the most important precondition.
The review of the literature (Sano, 1995) and the analysis of the features proposed by the similar available tools examined earlier, have driven the development of an interactive tool based on a subset and adaptation of these interaction techniques according to the typical expected user interactions.
In particular, having in mind to:

- ✓ allow for a drill down approach of 'overview first, zoom and filter, details-on demand';
- ✓ provide a user interface (UI) and a user experience (UX) as simple as possible;
- ✓ maximize the use of the set of data accessible and exposed by APIs;
- ✓ organize aesthetics (gird, axes, labels, colors, etc.) coherently with the dynamic features.

The theoretical visualization concepts discussed in this chapter underpins the methodological approach and the development of the visualizations and the web tool discussed in the next chapter.

4 ESIFy: a web tool for performance assessment

According to Shneiderman (Shneiderman, 1996) the best tools present information rapidly and allow for rapid user-controlled exploration. However, despite the substantial set of tools presented in the previous chapter, despite somehow they follow this agile approach, the depth and width of the information they cover is modest.

This chapter describes ESIFy[31], the web dashboard developed to enhance the use of open data for simpler and more effective interpretation and insights for policy making.

A dashboard is a simply presented complex composition of individual visualizations that have a coherence and thematic relationship between them (Few, 2006). This gives a holistic view of the phenomenon being assessed. Such compositions are widely used to analyze groups of variables and to support decision making, especially in private businesses.

This approach to data visualization improves two purposes for the graphical presentation of information, interpretation and communication, by combining a larger set of dimensions and measures related to the topic. This encourages further reasoning and analysis to support decision making. The objective of the dashboard is to allow a 'think twice' approach providing more food for deeper thought.

4.1 Visualize policy performance

The vast amount of information reported, data structure complexity and timing of updates means using agile tools to easily fetch, parse, aggregate and visualize information instantaneously, i.e. real time analytics.

Analysis of the data structure based on the EU reporting regulation could be based on the theoretical framework of the multidimensional model based on facts, events, dimensions and measures (Golfarelli, Maio, & Rizzi, 1998).

Table 1 – Dimensional fact model

Facts	Events	Dimensions (hierarchies)	Measures (units)
Project selection Expenditure declared Supported beneficiaries …	Project selection of OP_i, for TO_j in year *n* …	✓ Geo (EU, MS, OP) ✓ Time (year) ✓ Categorization (FUND, TO, PA[32])	✓ Allocated resources (EUR) ✓ Disbursed resources (EUR) ✓ Planned amount (EUR) ✓ Indicator achievement value (number, Km, EUR, etc.) ✓ Indicator target value (number, Km, EUR, etc.)

[31] https://eu-data.shinyapps.io/esify/

[32] There is a many-to-many relationship between Thematic Objective (TO) and Priority Axis (PA) with OPs.

The diverse sources hosting these data have common attributes that enable an associative model to create the data layer used in each visualization.

The proposed IT tool uses the R programming language, especially the *tidyverse* framework for data management and visualization and in particular the grammar of graphics defined in the *ggplot2* library (Wickham, 2017). Application back end relies on the *shiny* web framework for client-server interaction whereas the graphical and dynamic aesthetic of the front end interface is developed using the well-known Bootstrap framework and JavaScript (*plotly*). The web app is deployed in production in the shinyapps.io platform as a service (PaaS) for dedicated hosting of Shiny web apps.

As described in chapter 2, in order to maximize accessibility and exploitation by specialists and the general public, datasets are accessible and usable in different formats. In the ESIF portal, a web service ensures continuous access to data and metadata through API endpoints.

Data *and* the structure of data can be accessed through several endpoints in JSON format that allow fast fetching and parsing of microdata. The availability of semi-structured[33] data in the web service has driven the application back end architecture, requiring the development of specific fetching and parsing functions according to the data structure.

When the browser loads the application, a request for the JSON data objects is sent to the ESIF platform web service. Objects are parsed from the API endpoints exposing the ESIF database tables and metadata. These are then loaded in-memory on the server where database views (i.e. organized subsets of data for the specific output) are generated according to default parameters and then injected in the specific graphical object for rendering. The generation of view follows a dynamic approach according to new user inputs using simple tools as the dropdown lists.

Once data are loaded, there is short latency and high responsiveness for the application interacting with the user, especially for new plots.

The user inputs, single or in combinations, of Member State, fund and the name of the OP trigger a request to server for three main operations:

- ✓ Dataset association;
- ✓ Filtering and aggregation;
- ✓ Data visualization.

In particular, the subset of data is dynamically loaded and transformed in the processing algorithms to calculate the requested indicators. These (output) data are than processed

[33] JSON data are semi-structured data not organised in a relational model but with a clear hierarchical structure and organisation based on the *key:value* paradigm. In this sense, the JSON string provides both the microdata and the structure of the data.

by visualization algorithms with specific views in each of the main dashboard panels (e.g. EU level, Member State level, OP level, etc.).

Within each panel, the simple input menu allows searches for a specific Member State, Fund and OP but also to access the information of all other EU programmes for an immediate comparison even beyond the direct subject of interest for a specific user.

Even if the user selects more aggregation dimensions, visualizations are designed to always display all the aggregation levels, which avoids the zooming effect of displaying only the highest level of disaggregation.

As opposed to the state of the art and the tools described in previous chapters, this approach keeps the user experience as simple as possible but widens the range of accessible information to the maximum set of data exposed and available.

The dashboard design allows combinations of several visualizations at once in each panel and, for each of these views, displays several variables at once.

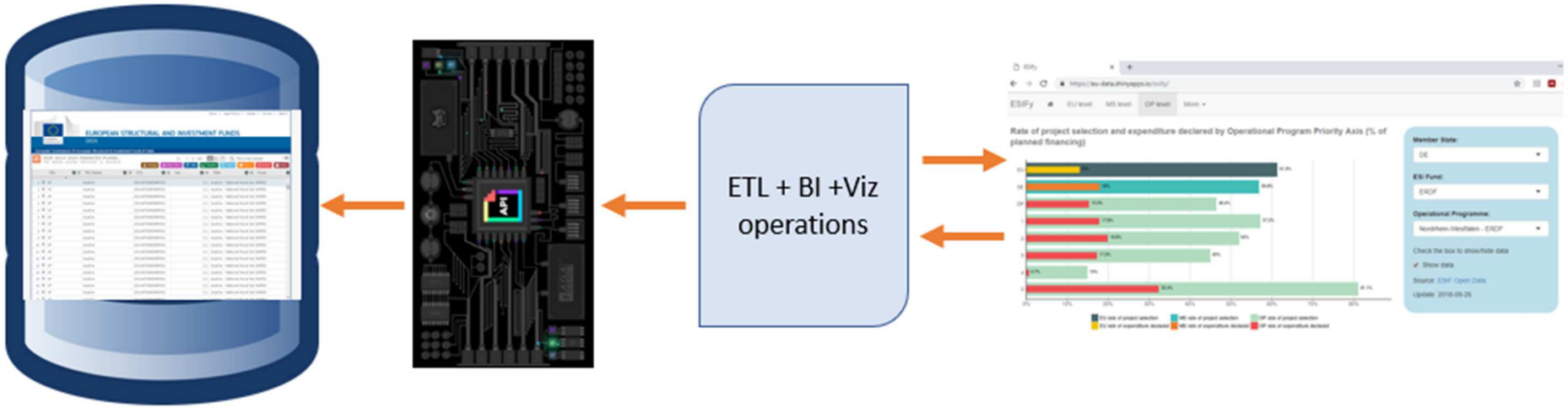
This implies an increase of information in terms of:

- ✓ number of figures;
- ✓ number of measures and dimensions per figure[34];
- ✓ number of units of observations (e.g. EU, Member State, regions) per figure.

Importantly, metadata relative to the data sources and update are included (using a dedicated endpoint) in the information available for the user. Although not always sufficiently considered and used, metadata are an important aspect in terms of trust, audit and usability of the information.

[34] According to the dimensional model, each aggregation/visualization has at least one dimension and one measure.

Figure 18 – ESIFy architecture



Source: own elaboration

Compared to the other tools described in chapter 3 such as the ESIF portal visualization or the ESIF viewer, the application is conceived as a **performance framework** interactive tool.

Data are presented in a performance perspective at the **deepest level available** (i.e. regional OPs) making use of all the information made available by the ESIF web service.

The main objective of data analysis and visualization for decision making is to easily compare regional programme performance to justify EU investment and inform taxpayers on the progress of deployed resources. So, all visualisations use a **benchmarking approach** either between different geographical levels (EU, Member State, OP) or over time, observing the progress since 2014.

Drill-down and benchmarking approaches have driven the design and aesthetic of the figures, with four specific **principles** considered when developing the visualization:

1. the **barplot** has been adopted being one of the most common, simple and recognizable types of data visualization, especially for policy makers and general public users;
2. the barplot has been improved through **nested progress bars** for the two main measures of resources, namely those allocated for selection and those disbursed for expenditure. This approach has three main advantages as it:
    - ✓ clearly highlights the strict dependence between the two variables;
    - ✓ immediately warns of anomalous progress (e.g. high ratio of selection and low ratio of expenditure);
    - ✓ excludes any risk of scale bias and misinterpretation;
3. as already discussed, **benchmarks** for space or time have been added to emphasize relative performance;
4. dimension-specific **patterns** for easier concept-insight association, i.e. OPs in descending order to focus on ranking; Priority Axis plot horizontally to compare advancements; Thematic Objectives as groups of bars to compare specific policy intervention fields.

The tool and the types of visualizations have received feedback and validation from a set of final users (UAT – user acceptance test) [35].

[35] The web tool prototype was presented at the EU datathon 2017 and during the European Week of Regions and Cities at the session *Open data in support of local and regional transparency, accountability, performance and beyond*, organised by the European Commission – DG Regio evaluation unit during October 2017.

4.2 Developing Key Performance Indicators (KPI)

A preliminary aggregation and cumulation of data at geographical and time level enables calculations that can be used to develop the performance indicators. These are mainly the ratios of planned allocations, namely Project Selection as a share of Planned Financing (EUR) and Expenditure Declared as a share of Planned Financing (EUR) disaggregated by specific dimensions.

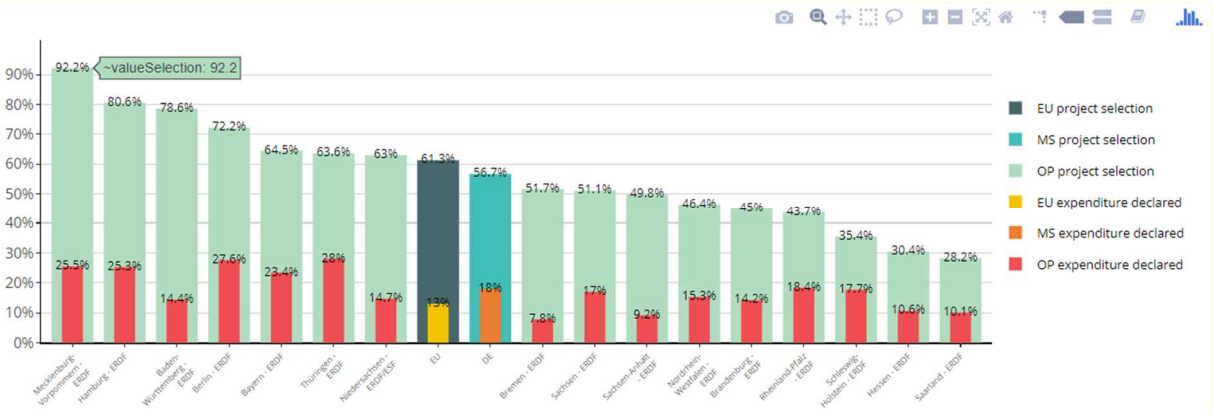
$$rate\ of\ selection = \frac{\sum Project\ selection}{\sum Planned\ amount}$$

$$rate\ of\ expenditure = \frac{\sum Expenditure\ declared}{\sum Planned\ amount}$$

The calculation and aggregation of indicators by itself does not give information to the user. For this reason, the visualization should be able to provide as much relevant information as possible without reducing a user's understanding of the message. Thus, a trade-off between the complexity of measures and dimensions has to be considered alongside the informative power when structuring a view. In this regard, depending on the specific information to transfer, each view aggregates and groups spatial or time dimensions as presented in the following figures.

Ranking OPs within the same Member State: ratio of selection and expenditure by OP, comparing all OPs in each Member State with reference to the EU level in decreasing order;

Figure 19 – Rate of project selection and expenditure declared by German OPs (share of planned financing)



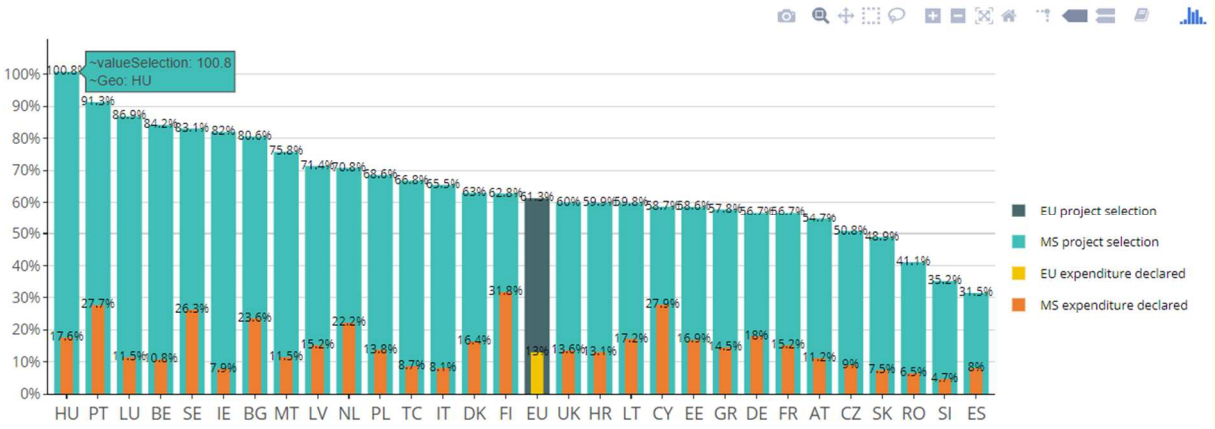
The algorithm behind the visualization ranks OPs within a specific Member State by rate of selection, as well as showing the Member State and EU averages. The figure shows that, despite the highest level of selection, the first two regions have lower levels of expenditure.

Even though some programmes show a very low level of progress, some German OPs are above the EU average for both selection and expenditure. However, overall implementation figure at EU level suggests **slow progress** with only a maximum of around 15% of resources disbursed at almost the half way point of the 2014-2020 programming period.

Ranking Member States: ratio of selection and expenditure by Member State, comparing Member States to the EU in decreasing order.

A similar approach could be followed at a higher level comparing all the 28 European Member States in terms of overall performance. This view zooms out from the Member State detail to provide general information on Member State progress in selection and expenditure.

Figure 20 – Rate of project selection and expenditure declared by Member States



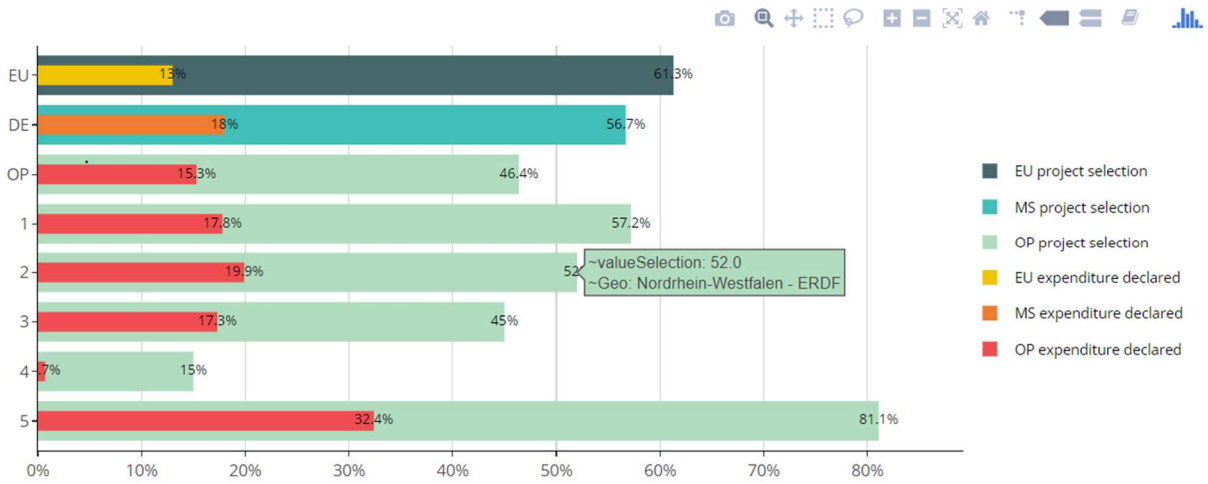
In the middle of the programming period, HU tops the ranking in terms of project selection whereas ES has only allocated one third of the planned resources.

However, FI proceeds faster as half the project financial resources have been disbursed.

This view synthetizes a large amount of data in a very important insights for policy making: the FI expenditure declared at around 50% of the selection suggest that compared to other EU countries it has the best **absorption rate**. This suggests that MAs of Finnish programmes are not doing better selection in general but better selection of **highly mature projects**, i.e. projects sufficiently well designed to shorten the time between the allocation and disbursement of resources.

Comparing Priority Axes (PA) of investment within OPs: ratio of selection and expenditure by PA, comparing all the axes in each OP to the overall OP, Member State and EU progress.

Figure 21 – OP rate of project selection and expenditure declared by PA (share of planned financing)



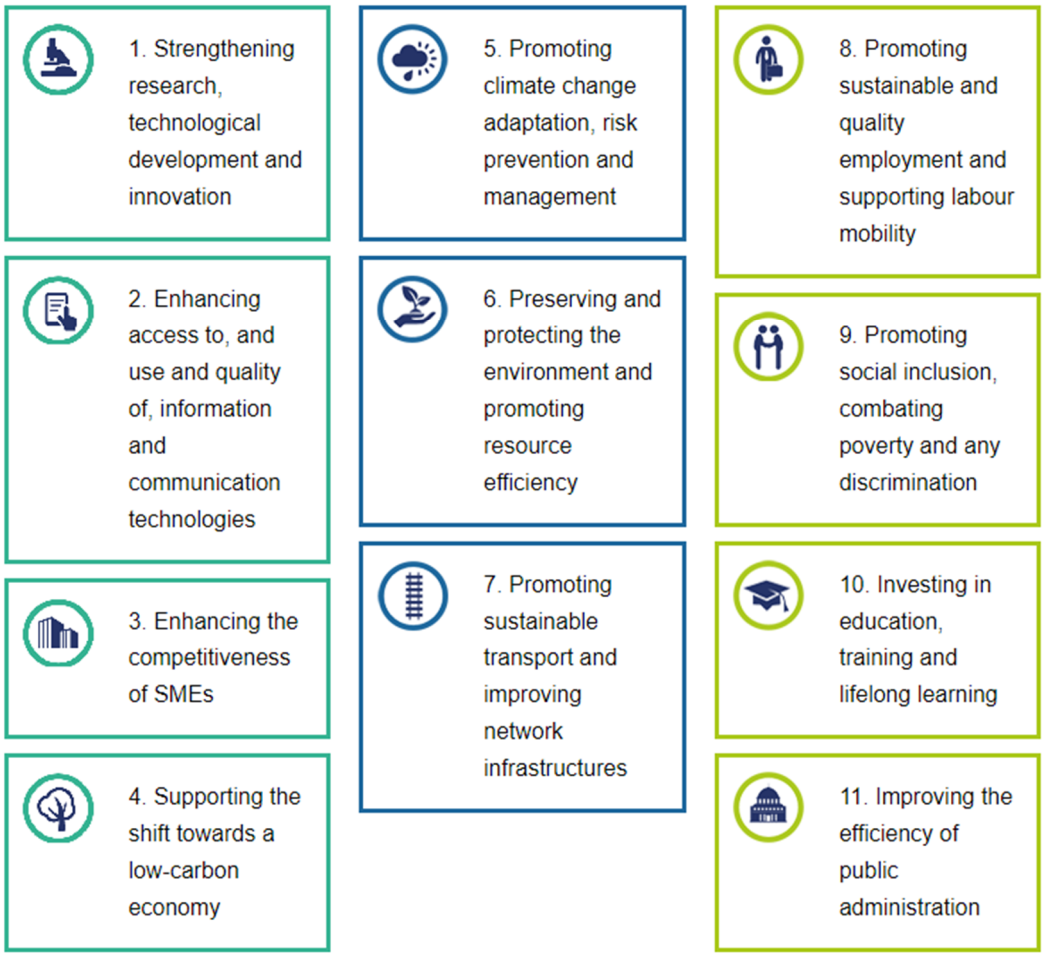
The figure shows progress of the overall OP under assessment almost in line with EU and Member State averages. Almost all PAs of the OP are progressing in terms of selection faster than the EU and Member State averages. However, the visualization highlights that Axis 4 has a very low selection rate and almost no expenditure, affecting overall performance of the OP. This warns of possible issues and obstacles in delivering resources for the specific type of investments related to the priority which implies **uneven progress** of the OP. Arguably, either low accessibility or scarce relevance for the possible applicants could be the reasons behind these patterns. In this context, accessibility refers to the selection criteria in the application phase being too restrictive and relevance to low interest of the measure for the target group.

Comparing areas of investment: ratio of selection and expenditure by TO with benchmarks to the Member State and EU.

The Europe 2020 strategy prioritizes Smart, Sustainable and Inclusive growth for European Member States. These priorities are than disaggregated and categorized in 11 TOs, which are key investment areas for a modern economic system.

The goal of these objectives is to concentrate financial resources on areas that deliver the highest benefits to citizens, fostering synergies between these fields and avoiding an excessive fragmentation of funding.

Figure 22 – The 11 TOs for the period 2014-2020

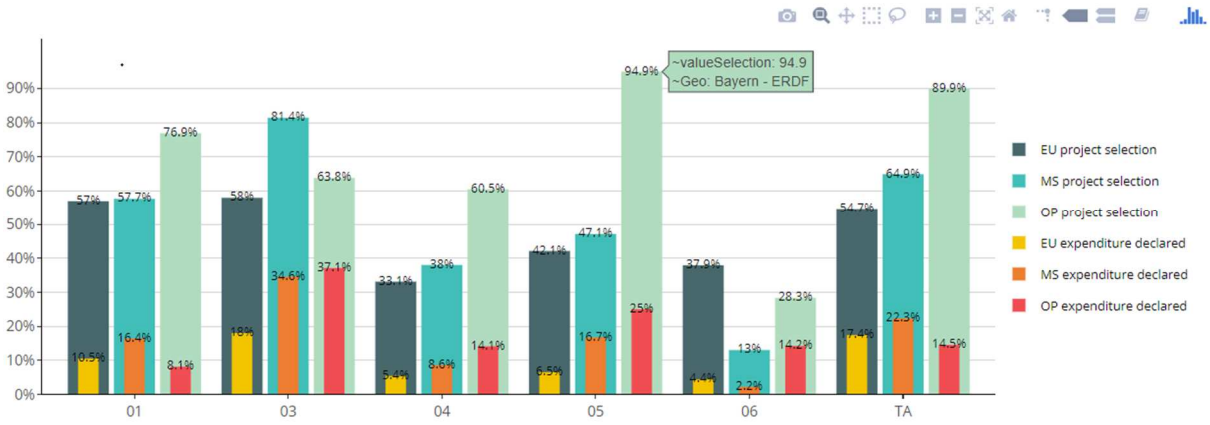


Source: European Commission

In terms of data management and visualization, TOs are simply another category for data aggregation and possibly combined with the other dimensions.

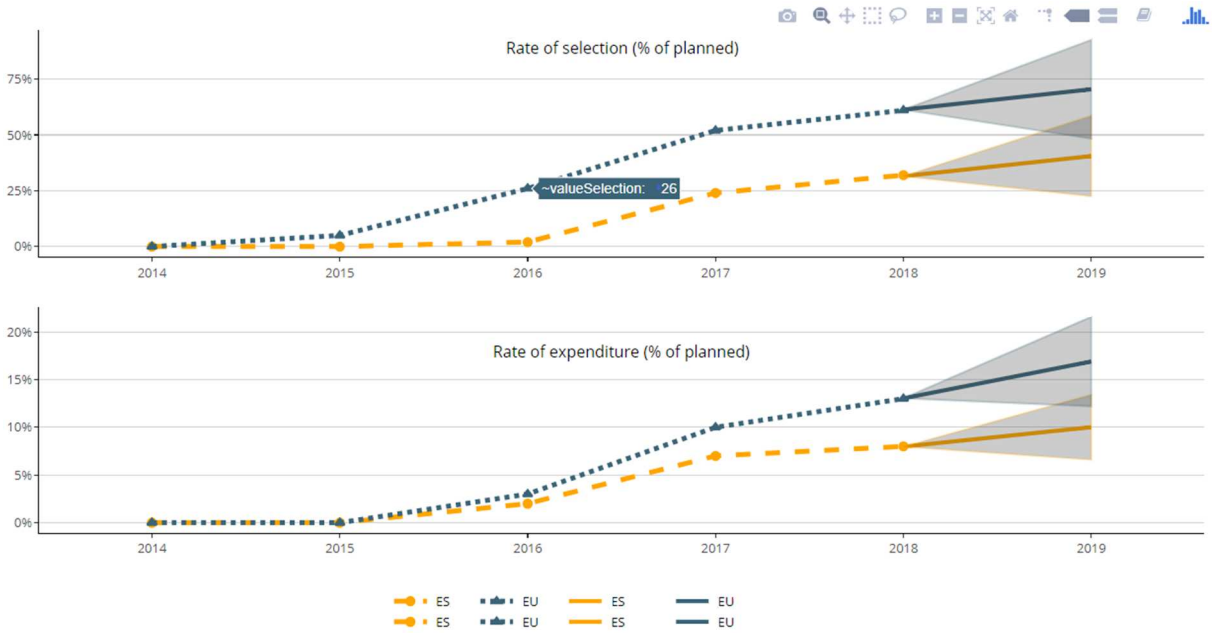
The following view allows comparison *within* TOs and *between* TOs. The figure shows the OP is performing slightly better both in terms of selection and expenditure compared to the Member State and EU levels in TO 4 and TO 5. Despite the highest level of selection in TO 1, expenditure declared is below the EU and MS level whereas the opposite occurs in TO 3. Comparing the different TOs of each level (comparison between) there seems to be lower progress difference in the EU level whereas MS ranges from an minimum of 13% (2%) to a maximum of 81% (35%) and OP level from 28% (8%) to 95% (37%).

Figure 23 –OP rate of project selection and expenditure declared by TO (share of planned financing)



Time series and prediction: time series of ratio of selection and expenditure over 2014-2019 with a year forward forecast and prediction band.

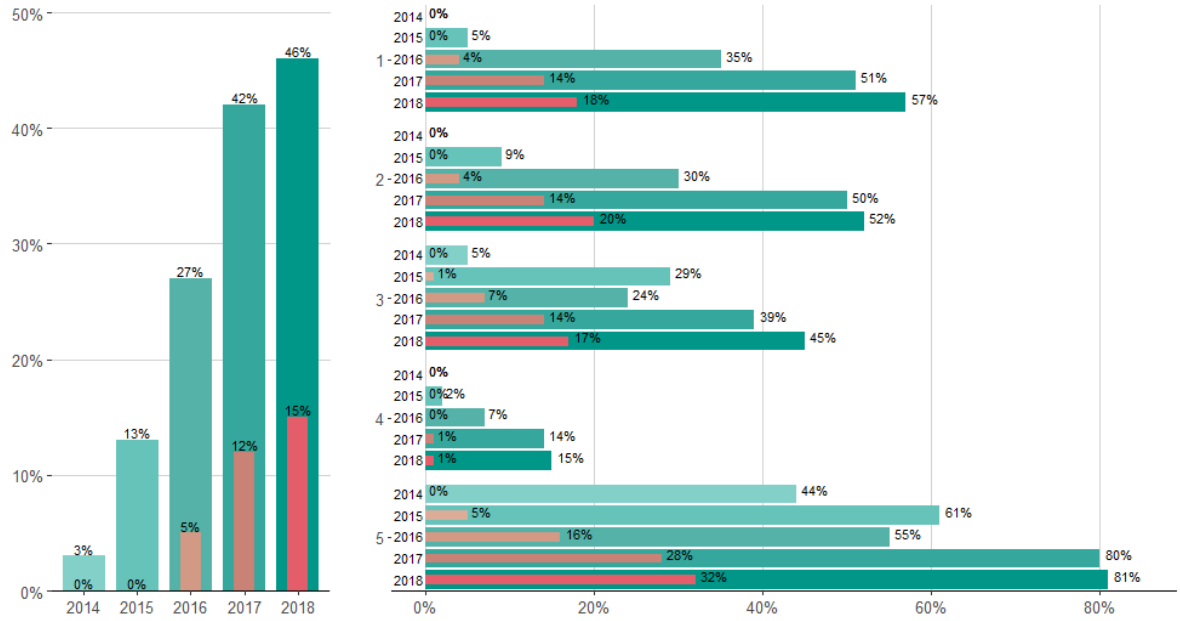
Figure 24 – Time series of rate of project selection and expenditure declared



In the period visualized, Spain always underperforms both in terms of selection and expenditure compared to the EU level, and an outlook with low probability of inverting the patter. The figure is also developed adding the OP time series and forecast.

Comparing progress over time: ratio of selection and expenditure by year between 2014 and 2018 with an overall visualization of the OP (left figure) progress and the development over time of the OP specific axes (right figure).

Figure 25 – Rate of project selection and expenditure declared over time: OP (left) and PA (right) details



Coherently with the natural progress of an OP over the programming period, the left-side figure shows the evolution of selection and expenditure with almost no resources allocated or disbursed in 2014 but an almost constant increase over the following years. The right-side view shows the progress over time of each PA with all developing in a similar way and, as already described in the previous figure, a critical situation in Axis 4. Here there was almost no change in selection or expenditure in the most two recent years.

Many other visualizations are organized within the application according to the menu bar on top of the dashboard as presented in the following figure.

There are three main panels:

- EU level with a general overview and comparison of the 28 Member States (Ranking Member States).
- Member State panel with details of all OPs (regions) within the Member State (e.g Ranking OPs within the Member State);
- OP panel with details of a specific OP against EU and Member State benchmarks (e.g. comparison of PAs, TOs and performance over time).

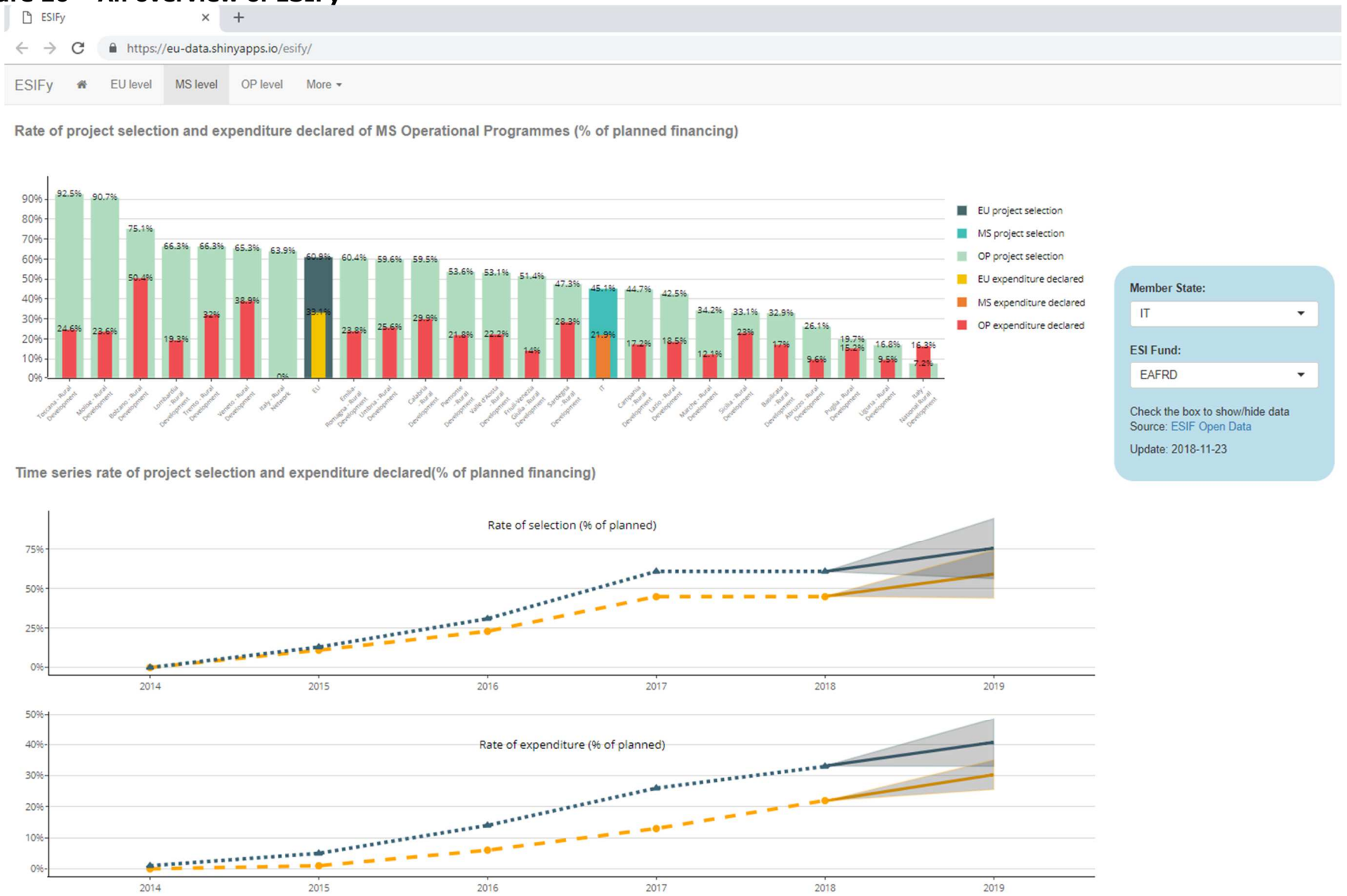
Additional sections (*More*) are dedicated to different datasets focused on measures and dimensions for:

- Output indicators;
- Financial Instruments.

The application full source code is available under GPL-3.0 in the Github folder https://github.com/rpalloni/ESIFy.

The main algorithms for data processing are in the *functions* folder divided into fetching and data analytics (i.e. ETL, BI and visualization operations). The source code creating each visualization organizes the three operations (Dataset association; Filtering and aggregation; Data visualization) within the same dedicated file.

Figure 26 – An overview of ESIFy



4.3 KPI for indicator achievement

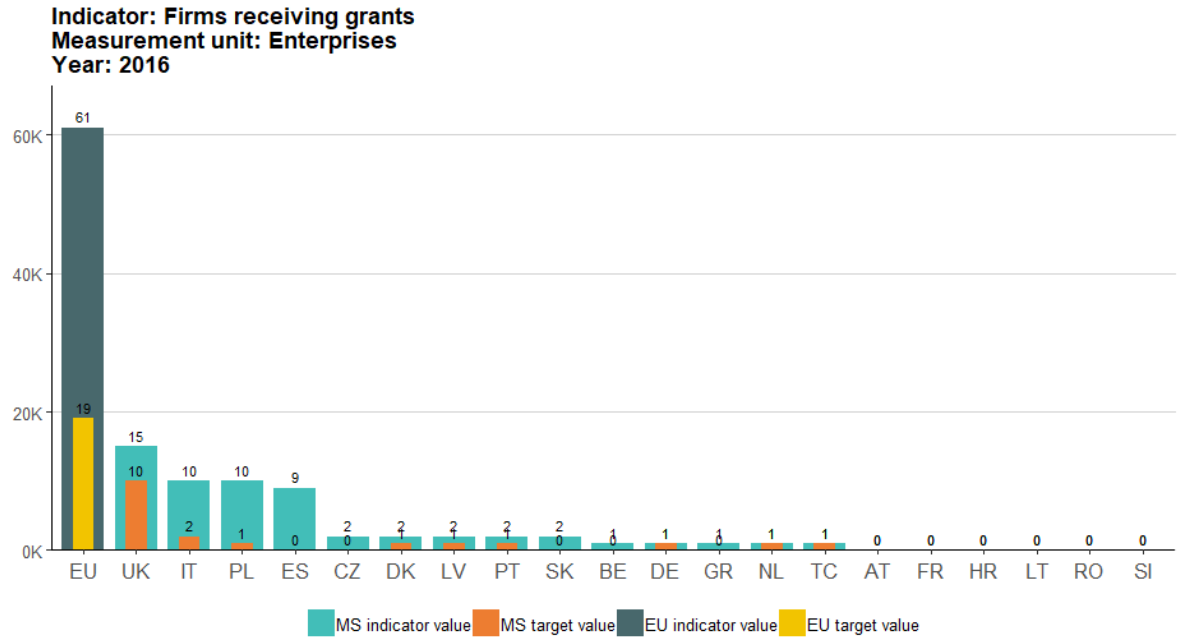
KPIs should not be limited to financial data as the tangible benefits for taxpayers and citizens in general are the effects generated by the expenditure of these resources.

For this reason, beyond the projects selected and expenditure declared, output and result indicators must be monitored by the MAs responsible for the management of funds and reported in the centralized information system in the framework of the AIR for European Commission approval. This reporting means these indicators are available as well as financial data in the open data portal through a dedicated API endpoint and categorized according to the dimensions required by regulation.

The core measures reported are the target values decided in the programming initial phase and the achieved values increasing annually.

Since each indicator monitors different elements, the unit of measure changes depending on the indicator.

Figure 27 – Nominal values of achieved and target values by Member State, CO02 (TO 01)

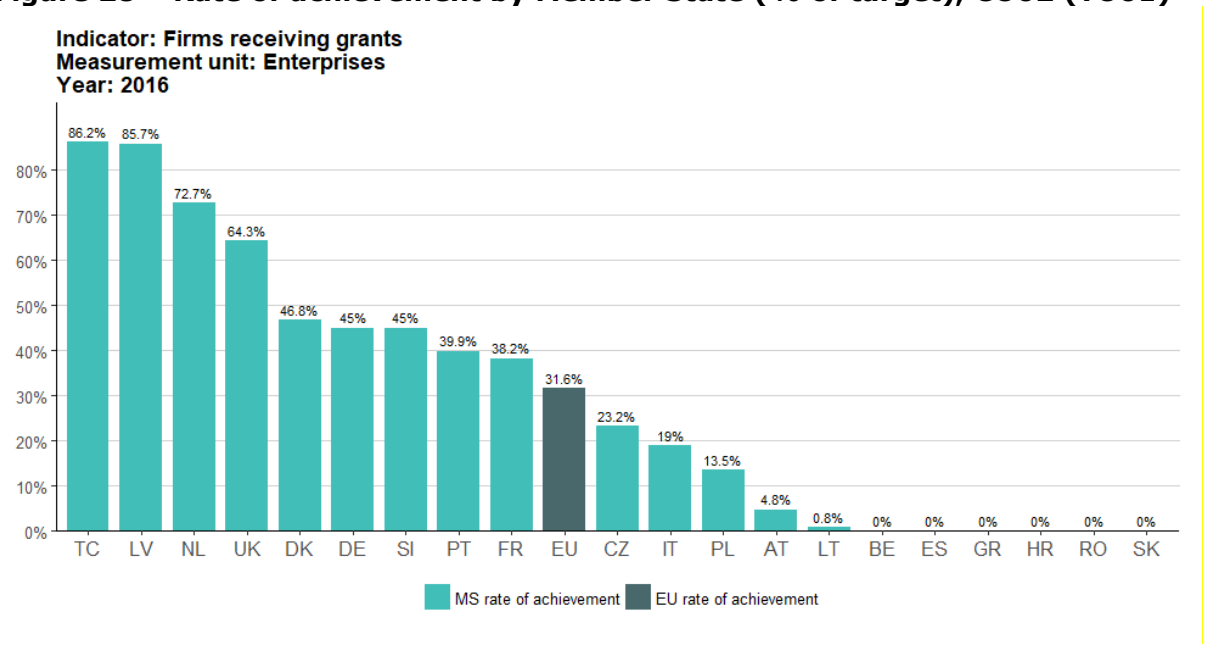


These two variables enable calculation of the new KPI rate of achievement with a similar approach to the rate of selection and expenditure as shown in the following figure.

$$rate\ of\ achievement = \frac{\sum Indicator}{\sum Target}$$

Ranking Member States: rate of achievement by Member State, comparing all Member States with reference to the EU level in decreasing order.

Figure 28 – Rate of achievement by Member State (% of target), CO02 (TO01)



Among the Member States reporting valid data, OPs dedicated to territorial cooperation (TC) have the highest rate of achievement followed by Latvia and the Netherlands. Some countries have not reported achievement values yet.

However, considering the reporting date, contrary to expectations the achievement rates are very high for the top group of countries suggesting target underestimation in the programming phase. Furthermore, the suspicion increases cross-checking TC financial performances as 35% project selection and 0.7% expenditure declared in 2016.

This is a clear example of the power of data visualization in spotting trends especially when combining the benchmarking approach to the multidimensional framework proposed in the dashboard.

Despite the possibility of linking insights between different measures, the target values of these physical indicators should be monitored not only compared to their achievements but also to the resources spent for each output unit.

For this reason, an attempt has also been made to define **efficiency KPIs**.

These could be considered as an approximation of efficiency and adopted for benchmarking. Selection efficiency is the ratio of resources allocated to projects compared to the amount in dedicated output indicators (e.g. number of firms receiving support, number of people employed, etc.). Similarly, expenditure efficiency divides the output by

the resources disbursed. In this way, two KPIs are defined based on million EUR spent per indicator unit.

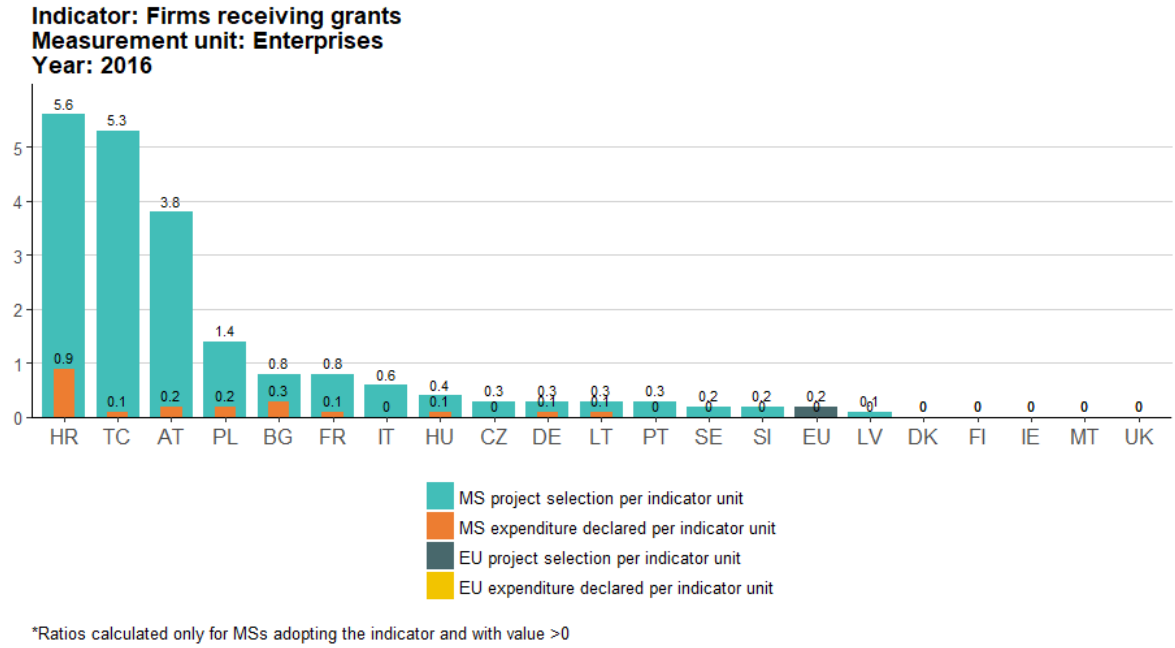
Again, benchmark-oriented visualization is developed based on disaggregation, using the many dimensions noted in previous sections (e.g. geo).

$$selection\ efficiency = \frac{\sum Project\ selection}{\sum Output\ Indicator}$$

$$expenditure\ efficiency = \frac{\sum Expenditure\ declared}{\sum Output\ Indicator}$$

However, calculating these indicators combining financial and indicator data reveals inconsistencies. The structure of the source data model between financial and indicator data is different due to the many-to-many relationship between financial aggregates of resources and indicators. The different levels of aggregation make it impossible to precisely match financial resources with the indicators. The resultant KPIs are only rough approximations of efficiency and a broad benchmarking comparison between Member States.

Figure 29 – Comparison of Member State selection and expenditure efficiency



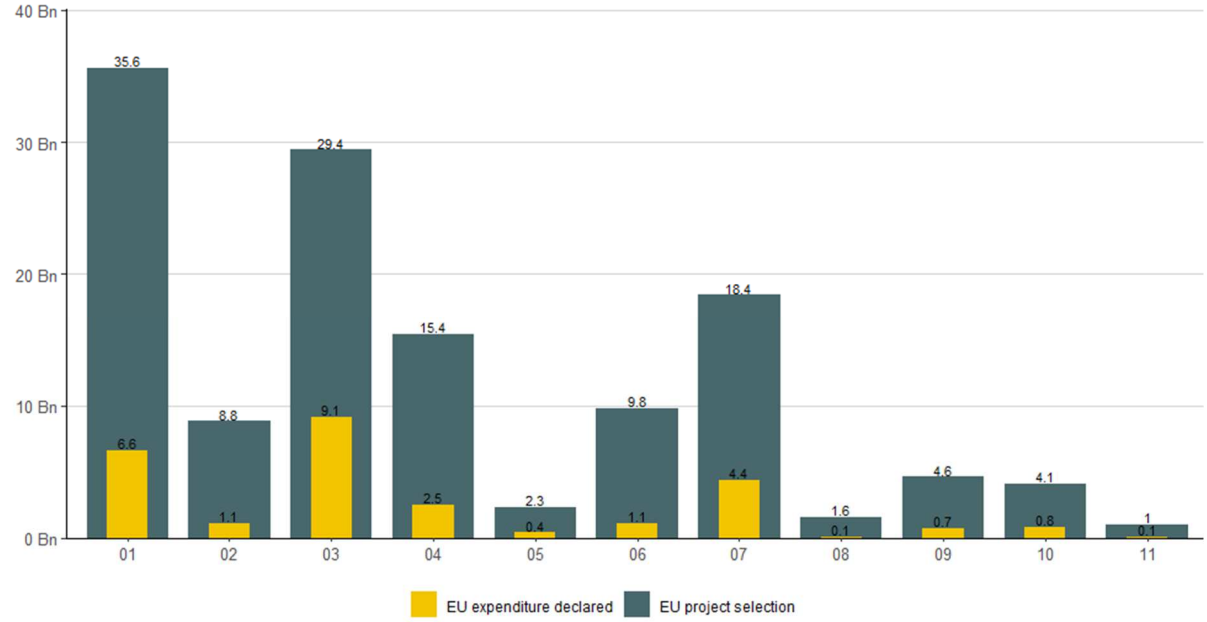
*Ratios calculated only for MSs adopting the indicator and with value >0

5 Open data for decision making

The 11 TOs to promote Smart, Sustainable and Inclusive Growth are presented in figure 20. The most strategically important for European economic development and considered as the modern industrial policy of the union is *TO 1 – Strengthening research, technological development and innovation*. The objective of TO 1 is to stimulate research and technological development in firms, especially SMEs.

The total budget dedicated to TO 1 for the 2014-2020 programming period is EUR 66 billion, mainly through ERDF, which is 10% of the total planned resources. The latest data update suggests that almost half the resources had been allocated to selected projects, though only about 10% or some EUR 6 billion had been disbursed.

Figure 30 – EU overview of TO project selection and declared expenditure

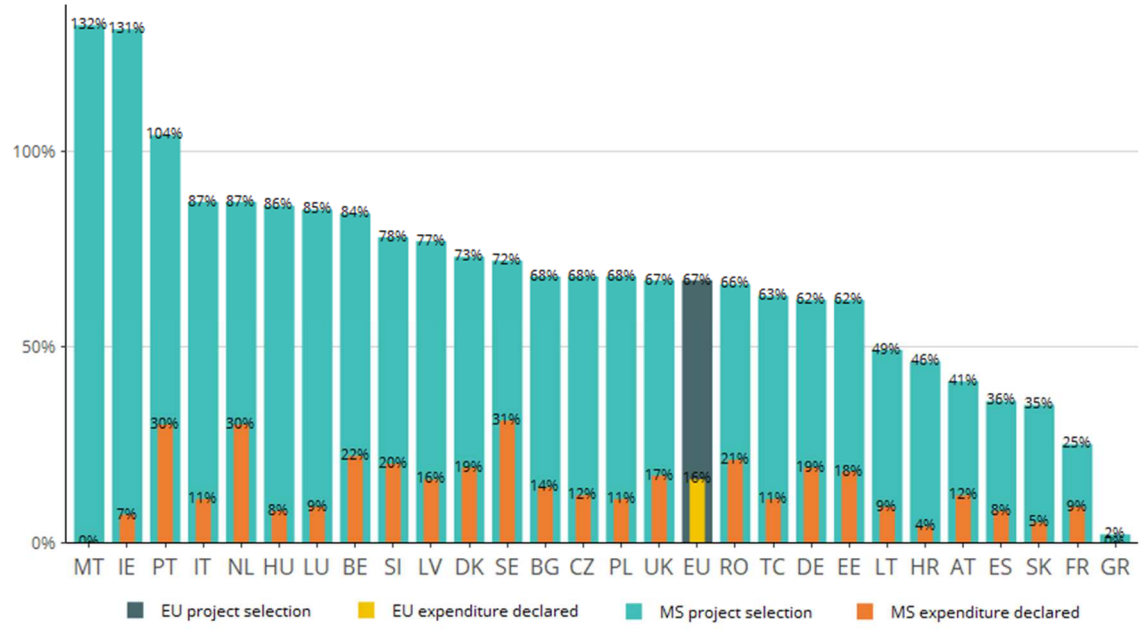


Since TO1 is the most important area for ESIF investment, this section focuses on using open data to extract insights and to highlight information relevant for managing financial resources dedicated to this crucial driver in European economic and regional development. According to Regulation, before planning and allocating investments under TO1, MAs must adopt a **Smart Specialisation Strategy** (S3) – also known as a Regional Innovation Strategy (RIS)[36]. S3 is a major new plank in EU innovation policy (Foray, 2015).

[36] According to the CPR, program authorities have a period to fulfil ex-ante conditionalities (including S3) in the first part of programme implementation. This means that even without the approved S3 the programme could be started and launched provided there is a plan to fulfil the obligation just after the approval. This was a compromise not to make the EC-regions negotiation on the S3 too much hindering for programme implementations.

Following the Lisbon strategy[37], the main aim of the 2014-2020 program is to foster innovative performance in EU regions and promote a better link between the production of new knowledge from R&D investment, and its application to new products and services. As shown in the following figure MS show uneven levels of selection and expenditure in TO1. So far innovator countries as Sweden, Netherland and Portugal have disbursed almost 30% of the planned amount, almost twice the European level.

Figure 31 – MS project selection and expenditure in TO1 (ERDF)



Within MSs, S3 means EU regions must identify the technological domains where they have superior innovative capabilities to be translated in competitive advantage in innovative products and services. S3 is designed so regions concentrate resources allocated to enterprises on projects in a limited number of technological domains. This should improve opportunities for the remarkable development of European research, innovation and technological diversification closer to citizens, i.e. at regional level.

The **choice** of domains should increase specialisation while also helping expansion into new sub-sectors. Furthermore, smart specialisation encourages strategic cross-border and trans-regional cooperation between regions to increase critical potential and the variety linked to that.

S3 combines recent theoretical advances in innovation policy and regional development (Foray, David, & Hall, 2009) as discussed in the following chapter.

[37] https://www.consilium.europa.eu/uedocs/cms_data/docs/pressdata/en/ec/00100-r1.en0.htm

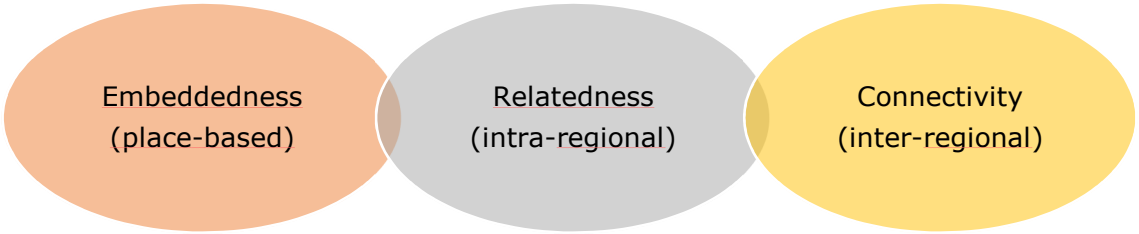
6 Theoretical background and literature

The S3 guide (Foray et al., 2012) emphasizes that regions should consider their specialisation domains in the design phase so the […] *S3 aims at developing world class excellence clusters and providing arenas for related variety and cross sectoral links which drive specialized technological diversification* (Foray et al., 2012,). Thus, the choice of technological domains is a real example of decision making that ought to be supported by data and in-depth analysis of regional dynamics.

The MAs choice of technological domains should theoretically be based on:

- Embeddedness;
- Relatedness;
- Connectivity.

Figure 32 – S3 principles



Source: own elaboration

6.1 Embeddedness

In particular, the S3 approach requires regions to select a few specialized **technological domains** on which to focus their innovation policy (Foray et al., 2012).

At the design level, the S3 approach has two main innovations (Foray, 2016). The first is that the domains should refer to technological domains rather than to industrial sectors (Asheim & Grillitsch, 2015; Foray, 2015).

This move from concentrating on industrial sectors to technological domains is the due to two theoretical arguments:

i) the relevance of R&D as a primary driver for innovation;

ii) the importance of knowledge capabilities for diversifying regional production.

The second novelty in S3 is an emphasis on the *bottom-up* approach in choosing the specialisation domains.

This choice of domains should result from an 'entrepreneurial discovery' process where actors in the regional innovation system, especially firms and universities, help identify the domains which should offer a competitive advantage. This help could be through their

participation in an analysis of the region's strengths and weaknesses to identify domains with the greatest potential for innovation and diversification.

The move from concentrating on industrial sectors highlights the increasing importance of investments in R&D and linking the results to innovation, as opposed to investments in applying existing knowledge to specific products and services.

Targeting technological domains instead of specific production is expected to enhance product innovation and diversification by creating new technologies and new forms of production (Asheim & Grillitsch, 2015; Foray, David, & Hall, 2011).

The emphasis on the bottom-up approach originates from new evolutionary economic geography (Ron Boschma & Frenken, 2011c; Lambooy & Boschma, 2001).

According to this theory, 'successful' regions can adapt and diversify their production in the face of changing conditions in markets and technologies. This process is path-dependent, meaning that it can be limited by past decisions and events which may no longer be relevant. The process also depends on the knowledge capabilities in the region (Neffke, Henning, & Boschma, 2011). This knowledge base can be seen in the technological know-how and organizational routines that can be applied to different products and services. Although products and services change rapidly according to market needs, the knowledge base does not as it comes from a long-term **accumulation process** (Balland & Rigby, 2017). Therefore, the ability of a region to diversify depends on its accumulated knowledge rather than on specific products or services. This is why innovation policy should target development of this knowledge base (technological domains) rather than specific products or services (Foray & Goenega, 2013).

Identifying and selecting technological domains in the region implies two important aspects in terms of strategy:

- Instead of spreading resources across many different fields, a region should concentrate on achieving critical mass and increasing the productivity of R&D investments (Foray et al., 2009).
- Regional renewal strongly depends on its industrial structure and infrastructure. As a result, regional specialisation is very path dependent making it important to consider the available innovative potential (Asheim, Boschma, & Cooke, 2011; Neffke et al., 2011).

However, identifying technological domains rather than industrial sectors involves two challenges, the lack of:

- suitable data and information about the knowledge base of a region.
- a shared methodology to identify promising technological domains for innovation and diversification.

Most of the data and information about firms and production are collected and organized under industry classification systems (NACE in Europe).

In theory, the 'entrepreneurial discovery' process should overcome both issues, at least partially. Firms, researchers and technology experts involved in designing the S3 should help regional governments by highlighting the most promising technological domains and research projects.

Firms and researchers directly involved in developing technology and products are better informed than policy makers about which domains are the most promising.

Regardless of the preference for a bottom-up or a top-down approach, the specialisation domains should be those with the highest existing or potential innovative capabilities in the region. Up to now, S3 documents contains only qualitative analysis about these issues, limiting any comparability of results within and between regions.

6.2 Relatedness

Evolutionary economics literature suggests that diversification into technologies 'related variety', can increase growth in regions (McCann & Ortega-Argilés, 2015). Such diversification within a region can benefit from knowledge spillovers and can encourage new combinations and relationships for industries (R. Boschma & Gianelle, 2013). Empirical studies confirm that related variety tends to encourage urban and regional employment growth (see e.g. Boschma and Iammarino, 2009; Frenken et al., 2007).

The concept of related variety has been important in the allocation of EU structural funds for the 2014-2020 programming period (McCann & Ortega-Argilés, 2015).

The emphasis on related variety in S3 policy is not surprising since technological relatedness has become a central concept in the literature about innovation and regional development. Related variety in a region should provide two main benefits. The first is to promote innovation through the cross fertilization of knowledge between different sectors (Ron Boschma & Frenken, 2011a; Frenken, Van Oort, & Verburg, 2007b). The second is to encourage diversification into new sectors (Ron Boschma & Frenken, 2011b; Neffke et al., 2011).

The related variety concept has been emphasized by several authors discussing the S3 rationale. McCann and Ortega-Argilés (2011) maintain that regions should specialize in different 'knowledge-related sectors'. This is relevant because: […] *domains which are highly connected with other domains will offer greater possibilities for learning than less connected domains* (McCann and Ortega-Argilés, 2011, p. 7). The authors underline that technological diversification through related variety is an important option especially in

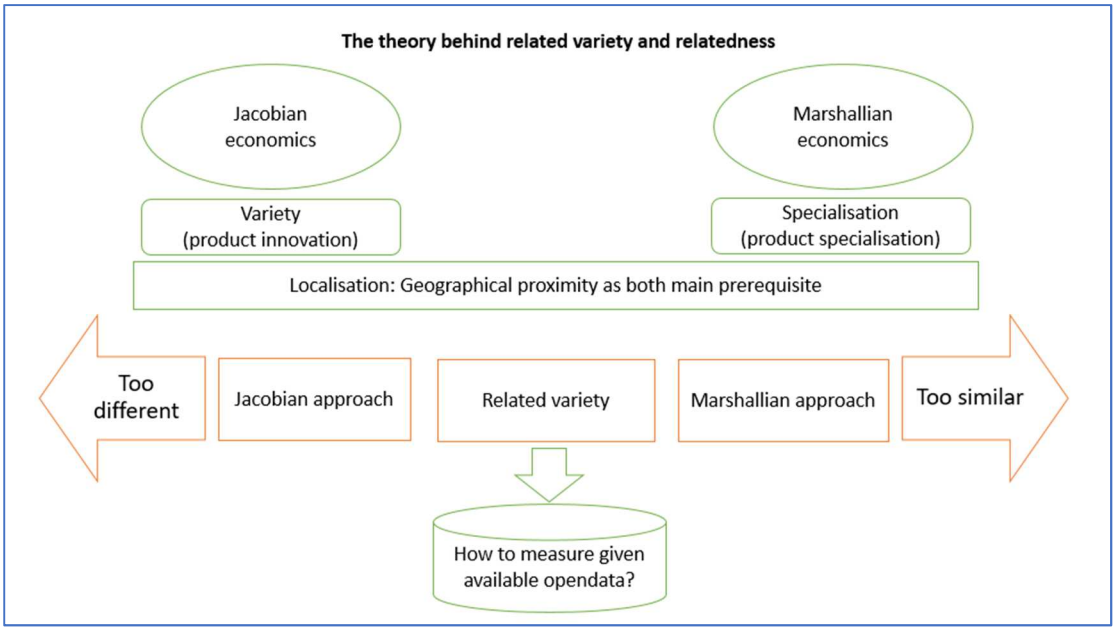
peripheral regions with excessive reliance on one or few technological domains. Boschma and Gianelle (2014) also agree that related variety between technological domains is not only beneficial to foster innovation performance and growth but also for diversifying the regional industrial base.

The concept emerged out of a long-standing debate on agglomeration economies and the different roles of specialisation and variety in promoting innovation (Duranton & Puga, 2001). Local cluster specialisation is expected to promote efficiency and incremental innovation but can hamper radical innovation and diversification as a lack of variety can result in cognitive lock-in, where the costs of change are overestimated (Ron Boschma & Frenken, 2011b, 2011c).

On the other hand, excessive diversity, through unrelated variety, can hamper innovation where effective communication and learning requires similar knowledge, or cognitive proximity (Ron Boschma, Eriksson, & Lindgren, 2009; Nooteboom, Van Haverbeke, Duysters, Gilsing, & van den Oord, 2007). Indeed, a recent strand of empirical literature demonstrated that it is not variety per se that matters but industries in a local area with complementary resources and knowledge (Ron Boschma & Iammarino, 2009; Ron Boschma, Minondo, & Navarro, 2010; Frenken et al., 2007a; van Oort, de Geus, & Dogaru, 2015).

The related variety approach could theoretically clash with the 'critical mass principle', which is a primary justification for the specialisation strategy. This problem is particularly important in small regions with difficulties in promoting multiple technological domains at the same time. The related variety approach is based on 'Jacobian' agglomeration advantages, where diversification and competition produce positive effects, though these are most often seen in rich and large urban contexts (Duranton & Puga, 2001; Jacobs, 1969).

Figure 33 – Related variety and relatedness



With this idea in mind, the EU guidelines for S3 design explicitly mention relatedness as one of the main criteria to consider when choosing specialisation domains (Foray et al., 2012). The S3 guide recognizes the importance of related variety as a driver for diversification, with the 'cross-fertilization' of ideas between different technological domains as a key factor to promote innovation, especially product innovation (Asheim et al., 2011; Frenken et al., 2007b; Grillitsch, Asheim, & Trippl, 2018).

In the last decade the concept of related variety has become increasingly important in the debate about innovation and regional development (Ron Boschma & Frenken, 2011a; Ron Boschma & Iammarino, 2009; Frenken et al., 2007a; Neffke et al., 2011).

Despite the importance of relatedness between specialisation domains, only a few regions explicitly considered the relatedness between technological domains as a criterion for their specialisation choices, according to an analysis of S3 documents approved by Italian regions (Iacobucci, 2014; Iacobucci & Guzzini, 2016b).

Applying the concept of 'related variety' in S3 design is not easy. As with the challenges for embeddedness, regional authorities do not have methodological indications for identifying technological domains to satisfy this principle, nor consistent sources of data.

This can be a weakness in implementing S3 since regional technological relatedness is considered a key factor for innovation and diversification (Asheim et al., 2011; Ron Boschma & Frenken, 2011a; Lambooy & Boschma, 2001; Neffke et al., 2011). Empirical evidence on the effects of S3 is still scarce (Caragliu & Del Bo, 2018).

6.3 Connectivity

The concept of connectivity proposed by the European Commission in the guide to S3 implementation looks for strong interaction between European regions in sharing and exchanging research and innovation.

[…] smart specialisation is pointing regions towards more strategic cross-border and trans-regional cooperation to achieve more critical potential and related variety (European Commission S3 guide).

In this respect, the connectivity concept supports and completes embeddedness, selecting technological domains closely related to the existing regional knowledge base, and relatedness, for the degree of connection between domains.

Beyond the requirements within the region, a strategic decision on selecting technological domains has to consider its position relative to other European regions. This implies looking for specialisation patterns beyond the regional administrative boundaries.

Regions should look for more strategic cross-border and trans-regional cooperation to ensure more potential and a related variety of research and innovation.

The S3 guide stresses the importance of identifying inter-regional links between technological domains, 'outward orientation', without offering a specific definition.

This 'outward orientation' should not rely excessively on nearby partners for learning and innovation, which could increase the risk of a region being tied to established industries (Hassink, 2005).

Fostering network relations with partners outside the region should avoid this.

It is also important to determine the physical distance of potential links. Geographic proximity helps with collaboration for innovation, as face-to-face relations are important for exchanging knowledge (Boschma, 2005).

The empirical literature also emphasizes the effects of geographical, technological and cultural proximity in establishing research collaboration (Cecere & Corrocher, 2015).

Moreover, labour mobility, which is a primary mechanism of knowledge exchange between firms, is greater in limited geographical areas. However, some authors argue that the role of geographical proximity in knowledge exchange and innovation is unclear, and that institutional, or social proximity for example may be more relevant. As a result, innovation networks increasingly rely on relations between firms and institutions in different regions and countries (Wagner & Leydesdorff, 2005). This means that, rather than being close, regions with potential knowledge links should have a similar knowledge base and institutional setting.

Moreover, as suggested by some authors (McCann & Ortega-Argilés, 2014), 'connectivity' with other regions may differ according to the degree of development in each regional innovation system.

This idea mainly covers potential links between core regions, at the frontier of technology, and peripheral regions, which use the technology for specific production (McCann, P., & Ortega-Argilés, 2013).

Peripheral regions should identify core regions with new knowledge that can encourage innovation in their specialisation domains. Regions emphasizing support for research rather than applications should identify regions with complementary research capabilities. The aim would be to encourage knowledge exchange typical in limited geographical areas. Within the core–periphery model, core regions could also identify regions that could apply the new knowledge.

These links would involve 'vertical relations' between producers and users of new knowledge, rather than the 'horizontal relations' implied by related variety. The benefits of cross-fertilization between sectors, which underlie the idea of related variety, depend on spatial proximity (Boschma, 2005) and are more likely with links within a region rather than between regions.

This analysis seems to reveal a contradiction in the S3 design. Under the S3 rationale, 'connectivity' within and between regions should be important in identifying specialisation domains and promising paths for diversification. At the same time, applying such concepts highlights several methodological problems hampering an analysis of connectivity within a region's S3.

Beyond justifying 'outward orientation' in S3 design, the S3 guide does not detail ways to identify existing and potential connections with other regions.

The lack of data and absence of clear methodology discouraged regions from attempting to analyse and measure connections with potential partners, even within the same Member State.

As pointed out by Iacobucci et al. *[…] Besides the rationale for justifying the 'outward orientation' in S3 design, little is said in the S3 guide about the* ***methods*** *that regions could follow to identify potential relations with other EU regions* (Iacobucci & Guzzini, 2016a).

7 Developing indicators

7.1 Methodology

In general, implementation of principles has been limited as there is no consolidated set of analytical tools for policy makers to use when designing and implementing their S3 (Balland, Boschma, Crespo, & Rigby, 2018).

One reason regions have not carried out an analysis is the lack of a consolidated methodology to measure ex ante and ex post the three aspects of embeddedness, relatedness and connectivity and the lack of data and information about how to apply the available methodologies.

Furthermore, the lack of quantitative approaches in S3 could be due to the natural, or non-standard, language used to identify technological domains. Regions have generally adopted a tree structure with two or three levels. The first level indicates the specialisation domains in relatively broad terms (i.e. 'health and wellness', 'mechatronics', 'aerospace', etc.). At this level, there are generally less than ten specialisation domains, even for large regions. The second and third levels specify the technological fields within the general domain, such as 'biorobotics for rehabilitation' within the 'life science' domain.

The use of **natural language** in defining specialisation domains (level 1) and technological fields (level 2 and 3) has several drawbacks. First, as largely discussed in information engineering, natural language owns the property of **ambiguity** (Anjali & P, 2014).

Second, the level of aggregation (or disaggregation) is not consistent across regions. Third, different labels may refer to the same technological domain. Third, the lack of common classification reduces comparability and hinders quantitative analysis of S3.

To overcome the limitations of this natural language, a homogeneous classification of technological domains is needed. The most obvious solution is the International Patent Classification (IPC). This is a knowledge categorization system designed by experts with several hierarchical levels that has been refined over 35 years.

In addition to the advantage of classifying technologies rather than products and services, using IPC enables the technological domains and knowledge base within the same region to be directly measured by the patenting activity.

Patents are a primary output of the innovation process. Moreover, they note the technological domains in which firms and research institutions are investing and accumulating knowledge. However, patents are not the only output of innovation and are unlikely to fully represent the knowledge base of a region. In some sectors, patents are less relevant for protecting technological knowledge, so their use is limited. In addition, while patents are a form of codified knowledge, most technological know-how is based on informal non-codified knowledge.

As previously discussed, one aim of the S3 is to promote more effective links between research and innovation. For this, S3 emphasizes the need to target technological domains where a region can develop new knowledge (with practical and economic relevance) by increasing R&D investment. The number of patents is a key indicator of the capacity to develop new knowledge in such domains. Moreover, patents are widely considered reliable proxies for the innovation performances at regional level (see e.g. Acs et al. 2002). Many empirical studies have based quantitative analysis of the innovation process on patents which are also considered a way to measure the value of knowledge over time.

However, the patent-oriented approach has been verified using innovation projects data of the CORDIS database for the 7th Framework Programme promoting research collaboration and involving different regional actors. Patents and projects are both outputs of the innovation process.

Another crucial aspect behind the difficult methodology development to measure the principles of S3 is the large difference in economic development between European regions, especially between the core and the periphery.

Several authors have questioned the relevance of S3 for less developed regions, as they may not have the knowledge base required to enable selection of the specialisation domains (Camagni & Capello, 2013; McCann & Ortega-Argilés, 2015). However, as a pillar of S3 is the pivotal role of firms[38], this may be even more relevant in less developed regions. Consequently, S3 in a less developed region should leverage any local innovation capacity in firms. Therefore, policy makers are expected to carefully consider existing local firm innovation capabilities when selecting the domains. From a methodological point of view, differences in innovative capacity between regions is taken into consideration using relative indicators of specialisation. Here, absolute specialisation indicators are more meaningful when applied to more developed regions with a steady stream of patents. On the contrary, relative indicators express the degree of domain specialisation with respect to other domains within the region. These are not affected by size, number of patents or level of regional development.

So, in less developed regions specialisation measures based on patents are good indicators of the accumulated knowledge base and local firms' innovation capacity.

Furthermore, the logic of specialisation for the S3 can (and in principle should) be applied irrespective of the level of technological development in regions. In theory, less developed regions may benefit from even more specialisation given their limited diversification

[38] For example, Foray (2015, p. 84) states that 'the centre of gravity of the smart specialisation dynamic is the firms since they are best placed to conduct entrepreneurial discovery processes. The strategy is much more broadly a tool for economic development through research and innovation that must associate all the actors concerned in projects not necessarily centred on public research or universities'.

possibilities. This avoids dispersing limited resources and concentrating efforts on domains where regions have nothing to build on and no chance of reaching critical mass. In other words, coherently choosing a limited number of technological domains may be even more important for less developed regions.

To relate the specialisation choices in S3 documents to the innovation capabilities of regions, the first step in **data** management has been the systematic association of technological domains noted in S3 documents (at the highest level of detail) with the corresponding IPC codes (at 3 digits).

7.2 Data

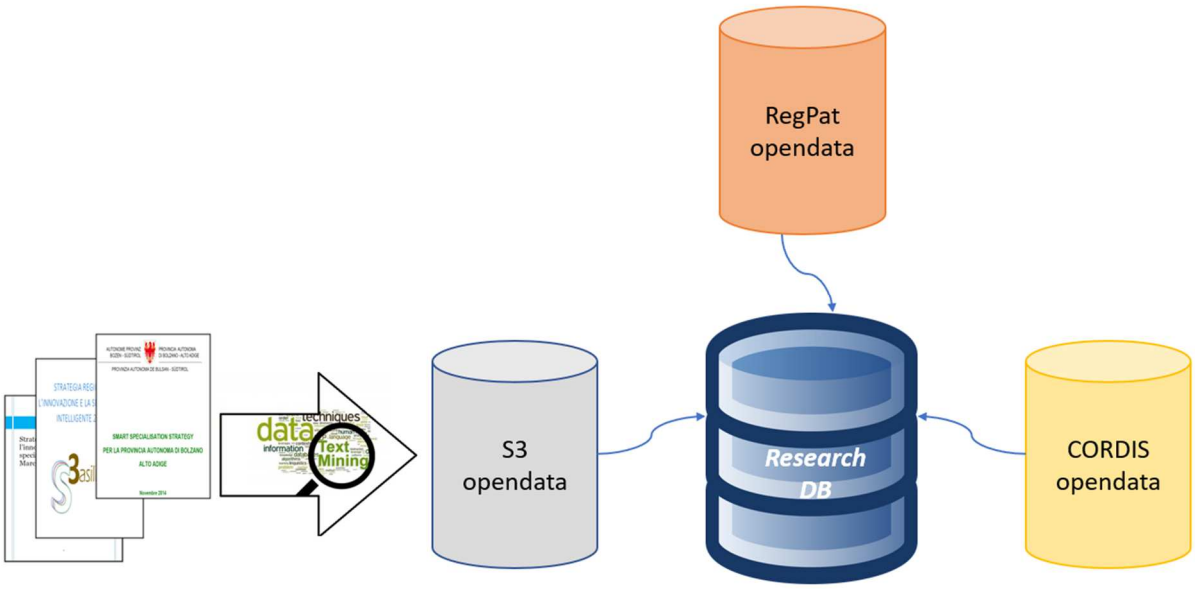
The sources of open data used in empirical analysis of the S3 principles are[39]:

- Smart Specialisation documents (definition of technological domains chapter);
- RegPat OECD database;
- Cordis Research and Innovation database.

The reference field to merge the information from each data source was the NUTS2 code, which is the key variable to model the relational structure of data between NUTS2 and IPC codes (one-to-many).

The analysis reviews 271 European regions and these are the unique reference for the data structure.

Figure 34 – Database for TO1 analysis: merging datasets



[39] Also Eurostat data via the web service endpoints have been merged to the core set of data as explained in the following (http://ec.europa.eu/eurostat/web/json-and-unicode-web-services)

S3 document data

S3 documents officially approved by Italian regions contain dedicated sections which explicitly indicate the investment domains. Although text mining could have been applied to all European regions for document data extraction, the use of national languages in S3 documents is a major obstacle, especially for data quality checking and manual refinement. However, Italy could be considered an ideal testbed for the methodology given the size of the country, the number of regions and their diversity in size, development and innovation. The following table presents reference indicators for Italian regions used to develop the methodology; population, density and innovation index.

Table 2 – Main indicators of Italian regions

Region	Population in 2016 (thousand inhabitants)	Rank pop. in 2016 (of EU regions)	Density in 2015 (inhabitants/ km^2)	Rank den. in 2015 (of EU regions)	Regional Innovation Index (RII) 2017	Rank RII 2017 (of EU regions)
Abruzzo	1327	179	122.7	165	66.19	132
Basilicata	574	277	57.1	269	59.42	146
Calabria	1971	110	129.7	157	59.29	148
Campania	5851	9	428.4	50	59.31	147
Emilia-Romagna	4448	19	198.2	111	81.98	98
Friuli-Venezia Giulia	1221	195	155.7	135	90.15	91
Lazio	5888	8	341.8	68	75.52	112
Liguria	1571	153	291.2	80	71.40	119
Lombardia	10008	3	419.3	51	81.65	101
Marche	1544	154	164.6	132	71.19	122
Molise	312	309	70.1	246	62.60	138
Piemonte	4404	21	173.9	127	81.85	99
P.A. Bolzano	521	284	70.2	244	71.24	121
P.A. Trento	538	282	86.6	216	80.42	103
Puglia	4077	24	209	108	60.07	144
Sardegna	1658	142	68.9	251	53.74	164
Sicilia	5074	13	196.8	112	52.70	168
Toscana	3744	30	163.1	133	77.46	107
Umbria	891	244	105.5	187	76.21	111
Valle d'Aosta	127	319	39.2	289	60.54	142
Veneto	4915	16	267.4	84	81.46	102

Sources: Eurostat and Regional Innovation Scoreboard, 2017.

The S3 documents were analysed to extract the technological domain chosen by regional authorities.

Since each region has defined these choices in a non-codified way, the first step was to homogenize the taxonomy to ensure fully comparable information. The second step was a systematic association between the most detailed description of technological domains in

the S3 documents and the corresponding IPC 3-digit codes. This association was semi-automated using the publicly available service IPCCAT (International Patent Classification Categorization Assistant). This service is based on machine learning algorithm of natural language processing (NPL) trained on manually classified documents to recognize IPC topics (Fall, Benzineb, & Guyot, 2003).

Then, the automatic mapping was revised and fixed manually.

For example, as shown in the following table, Lombardia had seven technological domains including Aerospace. Under these domains, the S3 document listed 13 specific technological sub-domains.

An IPC code was associated to each of them, resulting in a list of unique three-digit IPC codes per region. This enabled a detailed map of the technological domains and corresponding IPC codes.

Table 3 – Example of the semi-automated matching, domains-IPCs

Region	Regional needs and challenges	Technological domains	Corresponding IPC classes
Lombardia	• Ageing society • Health industry and wellness • Strengthening specialisation of the industrial and service system • Environmental sustainability • Digital divide and smart society • Improve mobility and accessibility	Advanced manufacturing	A23 A61 A62 A99 B01 B44 B60 B63 B64 B81 B82 C22 D01 G02 G06 G09 G21 H01 H02 H04 Y02
		Aerospace	
		Agrifood	
		Artistic and cultural industries	
		Green manufacturing	
		Health	
		Sustainable mobility	

The full list of regions and corresponding IPC classes is provided in Annex.

The technological domains were mapped into 64 IPC codes. The most commonly used code across regions was G06 (Data processing systems or methods). Of the 64 IPC codes, 20 (31%) were used once, suggesting moderate diversification for Italian regions in choosing their specialisation domains.

Each Italian region (k) has a set of IPC codes ($\mathbf{S_k}$) corresponding to the technological domains in its S3 strategy. In other words, if an IPC code (i) is in the set Sk then the region has indicated that technological domain in its S3 documents.

The use of IPC codes in characterizing specialisation domains is important for detecting similarities or differences in the domains. Indeed, some regions used the same (broad) label under which they combined different technologies or used different labels to refer to the same technological domain.

Patent data

Patent open data are available in the OECD RegPat database (February 2016 version; for a detailed description see Maraut, Dernis, Webb, Spiezia, & Guellec, 2008). This provides information about which IPC codes a patent belongs to and the address of its applicant(s) and inventor(s). Patent Cooperation Treaty (PCT) patent applications from 2002-2012 with at least one inventor based in Europe were each assigned a European NUTS2 region[40]. These data enable an assessment of the dynamics of innovation and technological specialisation within European regions.

It is important to note that 2012 was immediately before MAs started developing S3 strategies conceiving the document to be consistent with the foreseen regulation ex ante conditionality.

The RegPat database provides a share of PCT applications. If a patent was classified in more than one IPC class, its share is considered for each IPC class in each year and each European NUTS2 region.

CORDIS data

CORDIS is the Community Research and Development Information Service. It is the primary repository of EU-funded research projects and their results, including the formal deliverables.

CORDIS open data relative to organizations participating in research projects were used to validate the patent data. The results using RegPat data were checked against patterns of collaborations in projects between organizations in European regions.

There was an in-depth preliminary data cleaning and check by Member State using algorithms for cleaning procedures as proposed by Wickham (Wickham, 2014).

Furthermore, for the geographical dimension, although the database provides project organization postal codes, additional algorithms were developed to assign NUTS2 to each postal code using a two-step recursive procedure.

ETL operations on the datasets allowed obtaining structures of data for the calculation of NUTS2 level indicators that could be used to validate the three S3 criteria.

[40]Fractional count of PCT applications for each 3-digit IPC class: if a patent was classified in more than one IPC class, its fractional count is considered for every IPC class it belongs to

7.3 Measures of embeddedness

Empirical assessment of the coherence (embeddedness) was based on three patent-based measures using the IPC class:

i) relative specialisation;

ii) positive trend in the relative specialisation in the time window;

iii) absolute number of patents.

These measures were developed combining S3 document data and RegPat data to compare how congruent the technological domains chosen in S3 by Italian regions are and to highlight those with research and innovation capabilities, as measured by their patenting activity over time.

Relative specialisation measure

The Balassa Index is used to measure the relative technological specialisation at NUTS2 level, also known as the Revealed Comparative Advantage (RCA) index.

The RCA is defined as:

$$RCA_{k,i} = \frac{X_{k,i} \Big/ \sum_{k=1}^{K} X_{k,i}}{\sum_{i=1}^{I} X_{k,i} \Big/ \sum_{k=1,i=1}^{K,I} X_{k,i}}$$

where $\boldsymbol{X_{ki}}$ is the sum of the share (fractional count) of PCT patents for 2002-2012 in region $\boldsymbol{k}$ belonging to IPC class $\boldsymbol{i}$.

Therefore, $RCA_{k,i}$is the ratio between the patent share of region $\boldsymbol{k}$ in IPC class $\boldsymbol{i}$ and the patent share of IPC class $\boldsymbol{i}$ in the world. Since the Balassa index tends to have an asymmetric and skewed distribution, a symmetric version of this index applied the following non-linear transformation (Dalum, Laursen, & Villumsen, 1998):

$$RCA_{k,i}^{norm} = \frac{RCA_{k,i} - 1}{RCA_{k,i} + 1}$$

>0: Positive Specialisation → 1

<=0: Negative Specialisation → 0

The $RCA_{k,i}^{norm}$ is dichotomized using the threshold 0. Values below 0 indicate negative relative specialisation in a domain as identified by the 3-digit IPC class of a certain region, while values above 0 indicate a positive relative specialisation. This gives a binary $K \times N$ matrix (D_{bin}), where each cell is equal to 1 if the region $\boldsymbol{k}$ is positively specialized in the IPC class $\boldsymbol{i}$ and 0 otherwise.

Positive trend measure

As the innovation capacity evolves over time, a region could have significantly increased its technological specialisation in certain IPC classes between 2002 and 2012.

In this regard, it is worth noting that the concept of coherence is 'neutral'. It does not imply any judgement on the choices made by regional governments. They may have chosen domains with nor current specialisation neither promising trend because of future potential or after an entrepreneurial discovery process (promising domains). However, whatever criteria were used to include a specific technological domain, it is relevant to define its existing/evolving strength in the region.

This could show a promising specialisation to be considered, especially where the overall weight within the region is low, i.e. weak but growing.

To account for the time trend, a regression of $RCA_{k,i}^{norm}(t)$ is computed year by year, on the time variable $\boldsymbol{t}$ (i.e. year - 2002 + 1) for each combination of region $\boldsymbol{k}$ and IPC class $\boldsymbol{i}$.

A region $\boldsymbol{k}$ has specialisation growth in IPC class $\boldsymbol{i}$ when the coefficient of variable $\boldsymbol{t}$ is positive and significant (at the 80% level). Hence, the variable $trend_{k,i}$ is equal to 1 if region $\boldsymbol{k}$ has positive growth in the IPC class and 0 otherwise.

$$\beta \quad \begin{matrix} >0\text{: Positive Growth} \rightarrow 1 \\ <=0\text{: Negative Growth} \rightarrow 0 \end{matrix}$$

Absolute value measure

Finally, to have a measure of the absolute 'importance' of a region in an IPC class, the share of patents for each region for 2002-2012 is considered. The variable $NPat_{k,i}$ is the share of PCT applications of region $\boldsymbol{k}$ in IPC class $\boldsymbol{i}$ over the period 2002-2012. This measure identifies the number of patents with at least one inventor located in a specific European NUTS2 region. Then, the dichotomization also applies for this variable using the median of non-zero values across regions as the threshold.

7.4 Measures of relatedness

Empirical analysis of relatedness considers the intra-regional relations between IPC classes to develop indicators that can measure the intensity of their combination.

In particular, the methodology tries to detect the presence of relatedness indirectly on the basis of observed (revealed) associations between IPC classes: in each patent application may be detected by observing IPC codes within the same patents (Ponds, Van Oort, & Frenken, 2007).

Proximity measures

Intra-regional relations between IPC classes are quantitatively measured through the concept of 'proximity'. This underpins the methodology and construction of the relatedness measures in this document. Two measures are proposed:

- 'geographical / macro-level' measure based on Hidalgo et al. (2007);
- 'patent / micro-level' measure based on information on individual patents.

Jaffe (1986) uses this methodology to measure technological proximity.

Proximity à la Hidalgo based on relative specialisation measures

The relative technological specialisation at NUTS2 level for each region and class uses the calculation presented before to obtain the normalized Balassa index, i.e. the Revealed Comparative Advantage $RCA_{k,i}^{norm}$.

Following Hidalgo et al. (2007), the following measure of proximity is used for any pair of IPC classes *i* and *j* taking the minimum of the *pairwise conditional probability*:

$$proximity_{i,j} = min(\ P(RCA_i > 0 | RCA_j > 0), P(RCA_j > 0 | RCA_i > 0)\ )$$

where

$$P(RCA_i > 0 | RCA_j > 0) = \frac{P(RCA_i > 0 \wedge RCA_j > 0)}{P(RCA_j > 0)}$$

The proximity matrix is an $N \times N$ symmetric matrix with the revealed technological proximity between any two IPC classes *i* and *j*. Each cell (i,j) represents the probability that a region that is relatively specialized in *i* (*j*) is specialized in *j* (*i*) as well.

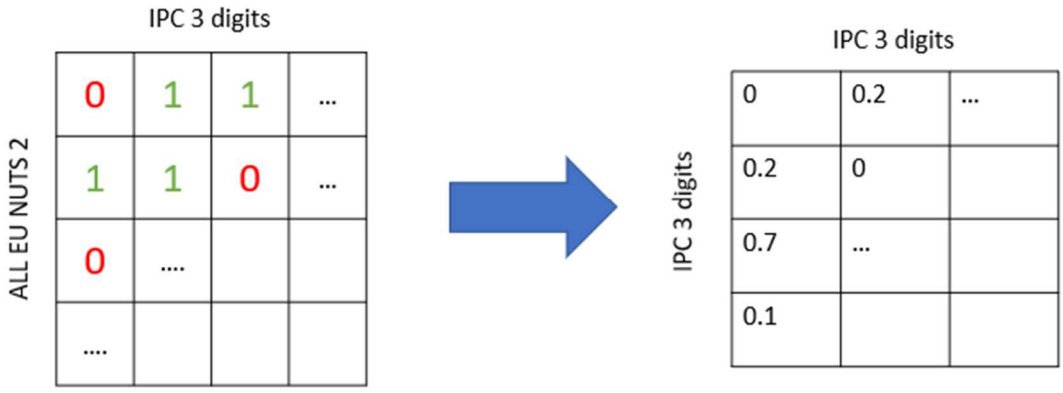
As an example, the cell (A01, A21) contains the probability value 0.4142.

This value is derived as follows:

1) Inner product of the $K \times 1$ RCA vectors A01 and A21. The product is equal to 58 (i.e. number of regions with co presence of specialization) and is divided by the total number of regions with specialisation in A21 (94 out of 270) resulting in 0.617.
2) Inner product of the $K \times 1$ RCA vectors A01 and A21. The product is equal to 58 (i.e. number of regions with co presence of specialization) and is divided by the total number of regions with specialisation in A01 (140 out of 270) resulting in 0.414.
3) Minimum between (1) and (2), i.e. 0.4142.

The following figure shows how the $K \times N$ matrix is used to create the $N \times N$ symmetric matrix containing the proximity measure for each combination of IPC classes.

Figure 35 – Proximity matrix in a nutshell



Therefore, this matrix is composed of the revealed proximities (probabilities) between technological domains as emerging from innovation activity in regions.

This matrix is symmetric and the diagonal elements are arbitrary assigned a value of 0. The idea is that when many regions are associated by specialisations in different technological classes, it should be optimal to combine (at the geographical level) specialisations to produce new technological knowledge (e.g. patents).

Proximity à la Hidalgo based on absolute value measures

In addition to the relative specialisation measure (Balassa index), the absolute 'importance' of a region in a certain IPC class is calculated with the share of patents for 2002-2012.

The variable $NPat_{k,i}$ is the share of PCT applications of region k in IPC class i during 2002-2012. This measure identifies the number of patents with at least one of its inventors in a specific European NUTS2 region. This variable is dichotomized using the first quartile of non-zero values across regions as a threshold, namely t. Use of this absolute measure is more in line with the idea of related variety since the exchange of knowledge between sectors depends on the resources invested in those sectors rather than their importance compared to other regions.

As with the approach for the relative measure presented earlier, a measure of proximity between any pair of IPC classes i and j is computed taking the minimum of the pairwise conditional probability:

$$proximity_{i,j} = min\big(P(NPat_i > t \mid NPat_j > t), P\big(NPat_j > t \,\big|\, NPat_i > t)\big)$$

where

$$P(NPat_i > t | NPat_j > t) = \frac{P(NPat_i > t \wedge NPat_j > t)}{P(NPat_j > t)}$$

The proximity matrix is an $N \times N$ symmetric matrix showing technological proximity between any two IPC classes i and j. Each cell (i,j) is the probability that a region

specialized (in absolute terms) in *i* (*j*) is specialized in *j* (*i*) as well. Therefore, this symmetrical matrix shows proximities between technological domains. The underlying idea is that when there is a frequent association (at the geographical level) between positive specialisations in different technological classes, it should be optimal to combine those specialisations to produce technological knowledge (e.g. patents).

7.5 Measures of connectivity

The connectivity measure is developed using a maximization approach. The idea is that a region interested in collaboration with other European regions would rather collaborate with a region that can maximize its innovation potential by combining their technological specialisations.

The operational approach to connectivity quantification follows methodologies and indicators to measure knowledge flows. The wide set of approaches adopted in literature range from patent citations (Jaffe et al., 1993; Almeida and Kogut, 1999; Maurseth and Verspagen, 2002; Fischer et al., 2006), to university collaborations (Katz, 1994) or co-publications (Hoekman et al., 2009; Scherngell and Hu, 2011), and research projects under the FP7 (Caloghirou et al., 2004; Breschi and Cusmano, 2004; Roediger-Schluga and Barber, 2006; Paier and scherngell, 2011; scherngell and barber, 2011).

To be consistent with the literature and to validate the approach based on patents, the empirical investigation on connectivity has two analysis steps. The first uses RegPat patent data to categorize technological domains in each region. The second is to explore and empirically assess measures of connectivity in terms of similarity and complementarity between European regions.

Similarity refers to the resemblance in terms of relative specialisation in technological domains between region pairs. **Complementarity** defines the degree of integration and specialisation potential when considering the pairs as a unique macro area.

Using this set of data clearly follows the approach adopted for the other S3 criteria.

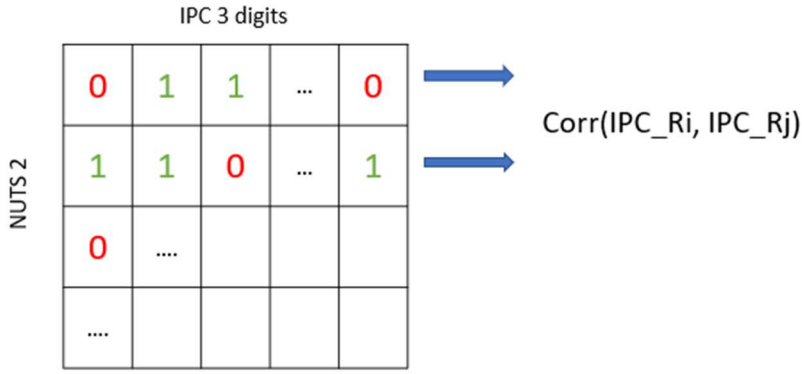
Another approach noted in the literature uses a collaboration measure derived from CORDIS research project data. In addition to developing a data-based measure, this also aims to validate the measure based on patents.

Similarity measure

The first measure developed is index of similarity which is the degree of resemblance between two regions. In statistical terms, it is the Pearson correlation between each pair of regions' specialisation vectors. The pairs of vectors are dichotomized so Tetrachoric

correlation has been used to verify the most common measure of correlation. The following figure provides an idea of the approach.

Figure 36 – Similarity matrix in a nutshell



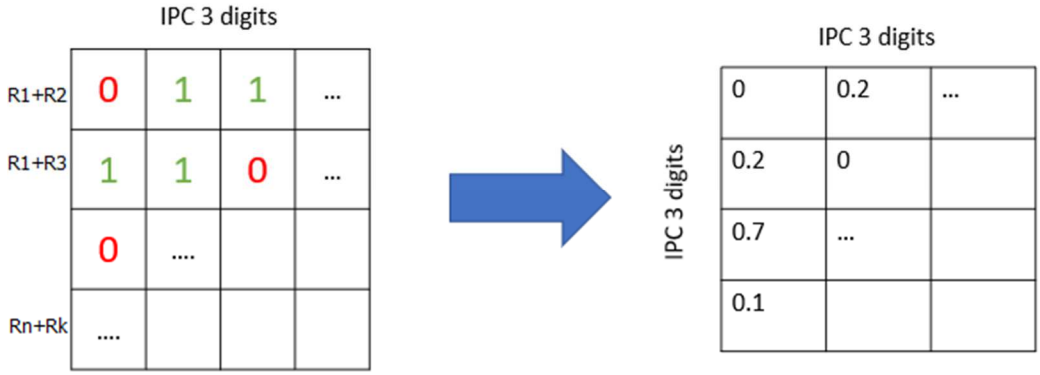
Complementarity measure

The second measure covers the specialisation integration between each pair of regions. In other words, complementarity measures relatedness in an 'augmented' region combining patents in each IPC class for each pair of regions.

The steps to develop the complementarity measure are:

1. Sum of patents for each IPC, for each pair of regions;
2. Calculation of the relative specialisation (Balassa index) in each IPC class for the 'augmented region';
3. Calculation of the proximity between technological domains;

Figure 37 – Complementarity matrix in a nutshell



Collaboration measure

Finally, CORDIS data for the 2007-2013 programming period were used to investigate actual collaborations in projects between organizations in European regions. More than 23 000 unique project IDs are available in the FP7 subset of projects with a maximum of 82 European regions in a single project (GRAPHENE).

However, most projects involved few organizations, with almost 60% covering up to three NUTS2.

On the basis of this data, a collaboration index for each pair of regions is calculated as the number of projects in which at least one participant was from region *r* and one participant was from region *s*.

8 Empirical results

The measures developed in the previous chapter are functional to developing indicators based on open data which are consulted for strategy development and assessment.

8.1 Embeddedness empirical results

The coherence between S3 innovation strategy and attributes in terms of innovative capacity is analysed for each Italian region through the following indices of relative specialisation, absolute specialisation and trend, as discussed earlier:

a. number of technological domains (IPC classes) noted in the S3 documents;
b. number of technological domains in which a region is specialized, as measured by the Balassa index, i.e. RCA;
c. **index of coherence** based on RCA: share of S3 technological domains in which a region is actually specialized, according to its normalized RCA index:

$$R_k = \frac{\sum_{i=1}^{N} I(i \in S_k) \cdot I\left(RCA_{k,i}^{norm} > 0\right)}{\sum_{i=1}^{N} I(i \in S_k)}$$

d. number of technological domains (IPC classes) in which a region is significantly increasing its specialisation (according to the positive trend measure over time)[41];
e. index of coherence based on RCA and positive trend measures: share of the chosen technological domains in which a region is:
   - ✓ either specialized (according to its normalized RCA index)
   - ✓ or significantly increasing its specialisation (positive trend)

$$T_k = \frac{\sum_{i=1}^{N} I(i \in S_k) \cdot I\left((RCA_{k,i}^{norm} > 0) \vee (trend_{k,i} = 1)\right)}{\sum_{i=1}^{N} I(i \in S_k)}$$

f. number of technological domains (IPC classes) in which a region has more patents than the median of non-zero regions;

[41] The indicators are based on current or foreseen specialisation in selected domains on the base of the evolution of patenting activity over time. However, the selection of domain could have been based nor on current neither on (recently) increasing specialisation.

g. coherence index based on the absolute value measure: share of S3 technological domains in which a region is actually relevant (the absolute number of patents within the top quartile at European level)

$$A_k = \frac{\sum_{i=1}^{N} I(i \in S_k) \cdot I(NPat_{k,i} > \underline{t})}{\sum_{i=1}^{N} I(i \in S_k)}$$

The coherence index could be clarified using an example. Region X is specialized (has a normalized RCA index greater than 0) in IPC class A01 (agriculture; forestry; animal husbandry; hunting; trapping; fishing) and A21 (baking; equipment for making or processing doughs; doughs for baking) and is not specialized in A22 (butchering; meat treatment; processing poultry or fish). Furthermore, the same region has declared in its S3 strategy that priority technological domains are A21 and A22 but not A01.

In this case, the regional strategy is 'coherent' for A21 and not coherent for A22, since it's not specialized in this technological domain. The resulting overall coherence index based on the relative specialisation measure will be 0.5 (1 domain in which the region is actually specialized of the 2 domains chosen).

As a real example of the coherence index, Marche region has 10 IPC classes associated with the technological domains declared in the S3 (see Annex) but only 3 are also among the 52 declared in the patent data from the RegPat database.

Figure 38 – An example of IPC revealed, declared and in common – Marche region

In general, regions have chosen about a third of the domains in which they show relative specialisation (i.e. RCA>0), i.e. 'owned but not chosen'. However, there are large differences around the mean. Toscana and Campania have a similar number of technological fields with relative specialisation (48 and 47 respectively): however, Toscana indicated only six technological domains in its S3 document, while Campania indicated 33.

Table 4 – Indicators of coherence

Region	Number of IPC codes in which the region is specialized (RCA>0)	Number of IPC codes chosen by the region in the S3	Index of coherence
Basilicata	29	13	0.31
Calabria	39	17	0.59
Campania	47	33	0.55
Emilia-Romagna	52	15	0.47
Friuli Venezia Giulia	43	16	0.44
Lazio	45	17	0.47
Liguria	50	16	0.50
Lombardia	70	19	0.42
Marche	52	10	0.30
Molise	25	5	0.40
Piemonte	59	15	0.47
Provincia autonoma di Bolzano	44	9	0.33
Provincia autonoma di Trento	48	22	0.50
Puglia	46	11	0.64
Sardegna	44	8	0.25
Sicilia	40	10	0.60
Toscana	48	6	0.17
Umbria	46	10	0.40
Valle d'Aosta	25	9	0.33
Veneto	56	13	0.62
Mean	*45.4*	*13.7*	*0.44*
Std deviation	*10.5*	*6.3*	*0.12*

Regions with a 'narrow' focus in their S3 (compared to their actual specialisation) are Marche, Sardegna, Molise, Provincia Autonoma di Bolzano, Umbria, Veneto and Sicilia.

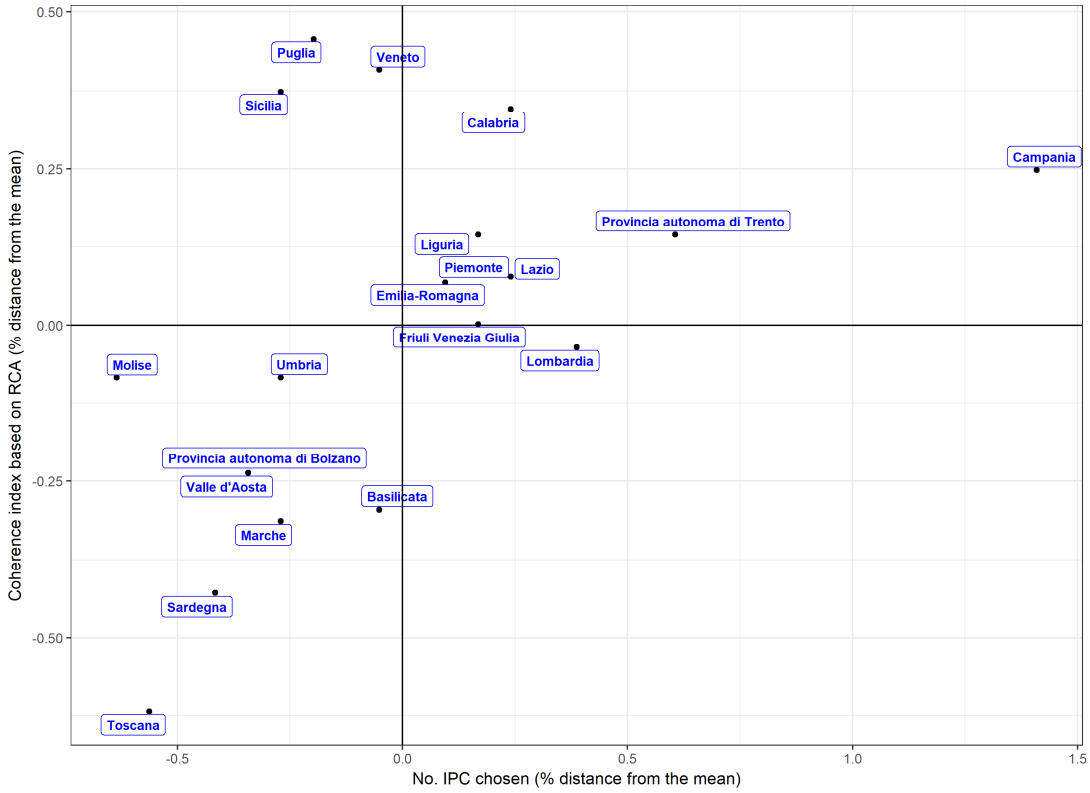
On the contrary, together with Campania other regions that opted for a larger span of specialisation are Basilicata, Calabria and Provincia Autonoma di Trento.

In addition to the breadth of specialisation chosen by regions, it is interesting to measure if the domains are where the region has a relative advantage (as measured by the RCA index). On average, the degree of 'coherence' is slightly less than 50%. Moreover, as with specialisation there are large variations around the mean, from Toscana and Sardegna at 0.17 and 0.25 respectively, to Puglia and Veneto at 0.64 and 0.62 respectively.

Using the span of specialisation and the degree of coherence, regions can be divided into four groups based on their distance from the mean of each indicator (see Figure below). In theory, the larger the specialisation span, the higher the risk of losing coherence in the choice of technological domains, which highlights a negative relationship between the two indicators. For this reason, regions should be distributed in the second and fourth quadrants. This is true for about half the regions. However, several regions have a narrow

span of specialisation but low coherence and a few regions that, despite a larger span of specialisation, have a level of coherence above the mean.[42]

Figure 39 – Regions by span of specialisation and degree of coherence (differences from the mean)



In addition to the index of relative specialisation (RCA), the analysis of potential technological domain coherence is based on two indicators.

The first is the trend in patenting to highlight technological domains where the RCA indicator increased before design of the S3. The second is an index of 'absolute' strength in terms of the number of patents per IPC class. A region may not have a relative specialisation in a technological domain but nevertheless its choice could be justified because it is increasing innovative activity in that area or has a critical mass of R&D activity and patents in it. This latter index is highly dependent on the size of the region and on its innovation capabilities. Large regions with a substantial and diversified knowledge base have more freedom in choosing their specialisation domains, as they have a critical mass of accumulated technological knowledge in many technological domains. This is not the case for small regions or regions with low innovation performance (most of the Southern regions in Italy) that had to choose from a much smaller set of promising technological

[42] The latter result is not surprising as regions chose only one third of the technological domains where they showed a relative specialisation.

domains. Table 5 shows the technological domains in which regions increased their relative specialisation before designing the S3.

Table 5 – Indicators of coherence between regional S3 technological domains and those in which the regions showed a positive trend

Region	Number of IPC classes with a positive trend	Share of S3 IPC where the region shows a positive trend	Share of S3 IPC where the region shows a relative specialisation OR a positive trend	Improvement from specialisation
Basilicata			0.31	
Calabria	2	0.06	0.59	
Campania	4	0.06	0.58	0.06
Emilia-Romagna	4	0.13	0.53	0.14
Friuli Venezia Giulia	4	0.06	0.50	0.14
Lazio	5	0.06	0.53	0.12
Liguria	2		0.50	
Lombardia	8		0.42	
Marche	6	0.10	0.40	0.33
Molise	1		0.40	
Piemonte	8	0.07	0.53	0.14
Provincia autonoma di Bolzano	2		0.33	
Provincia autonoma di Trento	5	0.05	0.50	
Puglia	4	0.09	0.64	
Sardegna	2		0.25	
Sicilia	4	0.10	0.60	
Toscana	4		0.17	
Umbria	4		0.40	
Valle d'Aosta			0.33	
Veneto	4		0.62	
Mean	*4*		*0.45*	
Std deviation	*1.8*		*0.13*	

There are far fewer than the number of technological domains in which the region is specialized. Moreover, of the 73 IPC classes in which the Italian regions increased their specialisation in 2002-2012 only 12 were included as specialisation domains in the S3. Half of these IPC classes were already considered in the previous coherence indicator, as they were already included in IPC classes where the regions showed relative specialisation. As a result, including the 'trend' indicator resulted in a significant increase in the coherence indicator for only a few regions.

The situation changes significantly when considering the absolute strength of regions in patenting, when they have more patents than the median for EU regions as shown in the

following table. As previously mentioned, the size of a region matters, as large regions have many technological domains in which they show absolute strength. On the other hand, only small or less-developed regions have no or only a few domains in which they show significant patenting activity. This produces a dichotomous distribution of the coherence indicator.

Table 6 – Share of IPC codes where region has patents near the EU median

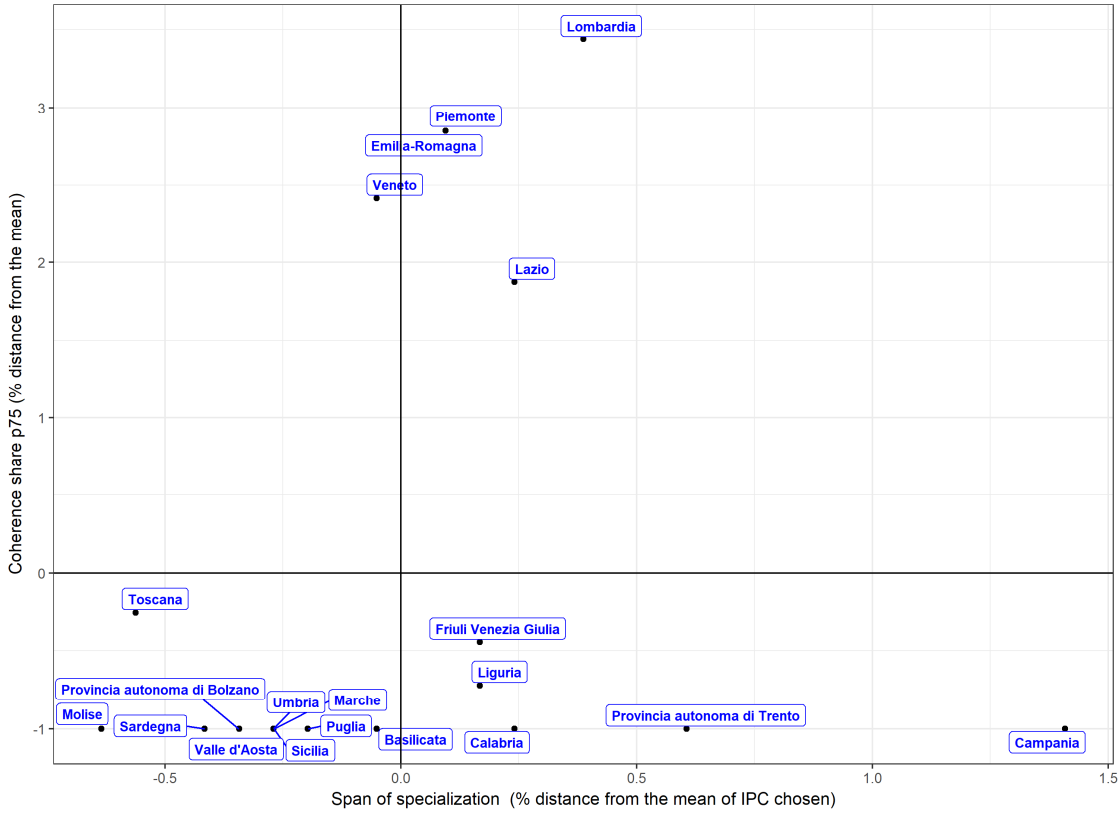
Region	Number of IPCs in which the region is over the European median	Number of IPC codes chosen by the region in the S3	Share
Basilicata	0	13	0.00
Calabria	2	17	0.06
Campania	28	33	0.36
Emilia-Romagna	109	15	1.00
Friuli Venezia Giulia	47	16	0.75
Lazio	87	17	1.00
Liguria	34	16	0.37
Lombardia	119	19	1.00
Marche	41	10	0.30
Molise	0	5	0.00
Piemonte	105	15	0.93
Provincia autonoma di Bolzano	6	9	0.00
Provincia autonoma di Trento	9	22	0.05
Puglia	13	11	0.27
Sardegna	2	8	0.00
Sicilia	5	10	0.30
Toscana	102	6	1.00
Umbria	7	10	0.00
Valle d'Aosta	2	9	0.00
Veneto	106	13	0.92
Mean	*41.2*	*13.7*	*0.42*
Std deviation	*42.8*	*6*	*0.40*

The largest and most-developed regions (such as Lombardy, Emilia Romagna, Veneto, Piemonte, etc.) have a coherence indicator of 1, as they could choose from a relatively large number of technological domains where they have significant patenting activity.
At the opposite end, small or less-developed regions (Basilicata, Calabria, Molise, Sardegna and Valle d'Aosta) had no technological domains with significant patenting activity.
Normally, there should be a positive relationship between the span of specialisation (i.e. the number of technological domains in which a region has chosen to specialize) and the share of IPC classes in which the region shows an absolute strength. The logic is that without an existing critical mass it would be more difficult to obtain a significant competitive

advantage in a domain. As a result, investment will be concentrated in fewer domains, as this will raise the likelihood of overcoming the initial weakness.

As shown in Figure below, this expectation is generally satisfied, but with some exceptions. Some regions have few domains in which they show absolute strength but have nonetheless chosen a large span of specialisation, resulting in a low level of the coherence indicator.

Figure 40 – Span of specialisation and share of IPC codes in which the region shows absolute strength



A robustness check, repeating the analysis using EPO applications instead of PCT applications, obtained similar results.

8.2 Relatedness empirical results

Based on the measure, several indices were designed to provide:

1. an ex-post measure of the degree of relatedness for the S3 specialisation domains;
2. an assessment of how much regions improved (with their S3 choices) their degree of relatedness as a result of technological specialisation.

For this, the following indices were computed for each region k:

$$\mathrm{ARI}_k = \frac{\sum_{i \in C_k} \sum_{j \in C_k, j \neq i} \mathrm{proximity}_{i,j}}{\sum_{i \in C_k} \sum_{j \in C_k, j \neq i} 1}$$

where $\text{proximity}_{i,j}$ is the proximity between the technological domains (i.e. IPC codes) *i* and *j* using one of the two measures proposed in the previous subsection, namely the proximity à la Hidalgo based on:

- ✓ relative specialisation,
- ✓ absolute specialisation.

The **Average Relatedness Index** (ARI), is an average of the proximities between any pairs of IPC codes belonging to $\mathbf{C_k}$, the group of technological domains chosen by region *k*. By way of example, if a region has chosen priority technological domains (IPC codes) A21, A22 and A01.

And in addition:

$\text{proximity}_{A21,A22} = \text{proximity}_{A22,A21} = 0.2,$

$\text{proximity}_{A21,A01} = \text{proximity}_{A01,A21} = 0.6,$

$\text{proximity}_{A22,A01} = \text{proximity}_{A01,A22} = 0.7.$

The index for the region is $\text{ARI}_k = \frac{0.2+0.6+0.7}{3} = 0.5.$

The **Relatedness Share Index** (RSI),

$$\text{RSI}_k = \frac{\sum_{i \in C_k} \sum_{j \in C_k, j \neq i} I(\text{proximity}_{i,j} > median)}{\sum_{i \in C_k} \sum_{j \in C_k, j \neq i} 1}$$

is the percentage of pairs of IPC codes chosen by each region *k* with proximity above a threshold. In particular, $I(\text{proximity}_{i,j} > median)$ is an indicator function counting when the proximity is above the median of all proximities.

Furthermore, the regional technological specialisation is compared with the optimal technological specialisation based on proximity.

In particular, for each region k two similar indices are calculated:

$$\text{ARI}_k^{actual} = \frac{\sum_{i \in S_k} \sum_{j \in S_k, j \neq i} \text{proximity}_{i,j}}{\sum_{i \in S_k} \sum_{j \in S_k, j \neq i} 1}$$

and

$$\text{RSI}_k^{actual} = \frac{\sum_{i \in S_k} \sum_{j \in S_k, j \neq i} I(\text{proximity}_{i,j} > median)}{\sum_{i \in S_k} \sum_{j \in S_k, j \neq i} 1}$$

ARI_k^{actual} is an average of the proximities between pairs of IPC codes in which region *k* is actually specialized. Specifically, $\mathbf{S_k}$ is the dichotomized vector of IPC classes in which region *k* is specialized (in relative or absolute terms).

Again, $\mathrm{proximity}_{i,j}$ can be measured as illustrated in the previous subsection, namely as the proximity à la Hidalgo based on 1) relative specialisation measures and 2) absolute specialisation measures.

As shown in the following table, all indices are negatively correlated with the number of chosen domains, although some of them only weakly.

Table 7 – Correlation matrix of relatedness indicators

		N° of chosen	*a*	*b*	*c*	*d*	*e*	*f*	*g*	*h*
a	**ARI$_r$**	-0.437	1.000							
b	**RSI$_r$**	-0.289	0.951	1.000						
c	**ARI$_a$**	-0.340	0.694	0.571	1.000					
d	**RSI$_a$**	-0.136	0.402	0.309	0.839	1.000				
e	**ARI$^{actual}_r$**	-0.122	0.139	0.090	0.186	0.175	1.000			
f	**RSI$^{actual}_r$**	-0.096	-0.049	-0.083	0.025	0.019	0.920	1.000		
g	**ARI$^{actual}_a$**	-0.014	-0.105	-0.052	-0.232	0.048	0.321	0.236	1.000	
h	**RSI$^{actual}_a$**	-0.112	-0.007	0.006	-0.124	0.099	0.329	0.184	0.892	1.000

This is not surprising given that the indices are an average of the all pairs of chosen IPCs. In principle, such an index can be maximized by taking the (unique) pair of IPCs with the highest proximity value. Adding any other IPC (less close to the initial pair by definition) would lower this index. There is a similar effect also when a region has two clusters with a high level of proximity within each cluster but a relatively low proximity between the two clusters. In this case the index would be lower than for a single cluster of related technological domains.

The correlation table also suggests a very high correlation between ARI and RSI, both using relative and absolute specialisation indices to measure proximity between technological domains.

The ARI is probably more precise while the RSI is easier to interpret; the high correlation between the two highlights that they are almost interchangeable for the empirical analysis.

Finally, there is a high correlation between the indices when using proximity measures based on relative and absolute specialisation. This is particularly true for the two ARI indices, which have a correlation close to 0.7. This means that the indices have low sensitivity to different formulations.

The table below shows the number of technological domains chosen by Italian regions and the ARI and the RSI indices computed as described in the previous section, using the proximity measures based on the relative and the absolute specialisation.

Table 8 – Relatedness indicators (chosen technological domains)

Region	Number of 3-digit S3 IPC codes (n_chosen)	Proximity à la Hidalgo based on *relative* specialisation measures			Proximity à la Hidalgo based on *absolute* specialisation measures		
		ARI_r	Fisher test p-value	RSL	ARI	Fisher test p-value	RSL
Basilicata	13	0.224	0.952	32%	0.690	0.159	74%
Calabria	17	0.397	0.000	90%	0.814	0.000	88%
Campania	33	0.276	0.354	49%	0.754	0.000	74%
Emilia-Romagna	15	0.277	0.391	46%	0.801	0.000	83%
Friuli Venezia Giulia	16	0.335	0.005	62%	0.749	0.011	66%
Lazio	17	0.284	0.270	46%	0.803	0.000	97%
Liguria	16	0.339	0.004	71%	0.783	0.001	78%
Lombardia	19	0.264	0.627	43%	0.740	0.012	69%
Marche	10	0.293	0.242	51%	0.713	0.135	51%
Molise	5	0.415	0.003	80%	0.865	0.001	100%
Piemonte	15	0.301	0.113	57%	0.755	0.012	72%
P.A. Bolzano	9	0.322	0.070	53%	0.792	0.006	83%
P.A. Trento	22	0.284	0.247	54%	0.688	0.088	66%
Puglia	11	0.275	0.439	45%	0.703	0.146	45%
Sardegna	8	0.298	0.216	54%	0.793	0.009	86%
Sicilia	10	0.328	0.041	64%	0.766	0.022	80%
Toscana	6	0.385	0.009	80%	0.783	0.036	67%
Umbria	10	0.345	0.012	60%	0.852	0.000	100%
Valle d'Aosta	9	0.363	0.006	64%	0.786	0.010	78%
Veneto	13	0.346	0.005	65%	0.785	0.002	85%
mean	*13.70*	*0.32*		*58.33%*	*0.77*		*77.12%*
std.dev.	*6.35*	*0.05*		*0.14*	*0.05*		*0.14*

As said, the number of technological domains chosen within the S3 is quite different across regions, with an average of 13 but ranging from 5 to 33 IPC classes.

This is not surprising given the large differences in the size of Italian regions, from less than 200 000 people in Valle d'Aosta to about 10 million in Lombardy.

The mean value of the ARI_r for Italian regions (0.32) is above the average proximity between technological domains (0.26) in EU regions (i.e. the expected proximity if two domains were chosen at random).

To provide a measure of the ability of regions to choose related sectors, a Fisher's exact test[43] was used to understand the 'distance' between the ARI of the regional choices and one resulting from a random choice of technological domains. Specifically, the p-value is the probability that a random extraction of IPC (from those in the region) would result in an ARI at least equal to the observed one giving an indication of the 'strength' of the choice with respect to a random selection.

The test highlights large differences in the behavior of regions. Some of them, like Calabria, Friuli Venezia Giulia, Liguria, Molise, Tuscany and Veneto, could significantly improve the ARI_r value compared to a random choice.

In other regions, the test reveals that the relatedness index value is not statistically different from the value that resulted when regional authorities selected technological domains in their region at random. The value of the test is logically related to the value of RSI, i.e. the share of chosen domains with greater technological relatedness.

In interpreting these results, it is worthwhile considering that ARI is negatively related to the number of chosen domains. This is because the more technological domains chosen by a region, the closer the ARI is to its mean value.

In other words, it is easier to have a high level of relatedness by choosing a few, highly related domains. With more domains, some must be less related to the others.

The situation improves significantly, both in terms of ARI and RSI, when using the indices based on absolute rather than relative specialisation. As mentioned in the previous section, the absolute specialisation is more logically connected with the concept of related variety at local level: i.e. the possibility of exchanging knowledge and resources between different technological domains.

Also, in this case (as previously observed for ARI_r), the mean of the ARI_a in Italian regions (0.72) is above the average proximity between technological domains in EU regions (0.68). Even more important, with a few exceptions, almost all regions chose technological domains with more relatedness than a random choice.

However, there may be a trade-off between maximizing relatedness (i.e. choosing domains with a high degree of relatedness) and maximizing coherence even though both these criteria were requested by the S3 logic.

For this reason, an alternative relatedness indicator is computed in the hypotheses. This assumes that each region had chosen the IPC codes of technological domains in which it is

[43] In particular, 10,000 random drawings of *n* IPC classes were performed, with *n* being the number of IPC codes chosen by the region in the S3. For each drawing, ARI is computed and the distribution compared against the ARI actually observed (the choices actually made by regions). The reported p-value is the probability that a random extraction of IPC would result in an ARI equal or greater than the observed one. Please note that the higher *n*, the lower the variance of the distribution of the simulated ARI, and the narrower the distribution close to the mean of proximities.

actually specialized. The *actual* ARI and the *actual* RSI are still computed using proximity measures based on relative and absolute specialisation, but consider actual specialisation rather than the choices made by regional authorities for S3. These indices give an indication of the degree of relatedness of technological domains in which a region is relatively or absolutely specialized.

Table 9 - Relatedness indicators (actual technological domains)

Region	Number of 3-digit S3 IPC codes (n)	Proximity à la Hidalgo based on *relative* specialisation measures		Proximity à la Hidalgo based on *absolute* specialisation measures	
		ARI^{actual}	**RSI^{actual}**	**ARI^{actual}**	**RSI^{actual}**
Basilicata	13	0.292	61%	0.823	100%
Calabria	17	0.335	72%	0.776	77%
Campania	33	0.304	61%	0.717	69%
Emilia-Romagna	15	0.345	84%	0.667	57%
Friuli Venezia Giulia	16	0.336	79%	0.718	71%
Lazio	17	0.316	62%	0.686	60%
Liguria	16	0.296	61%	0.710	66%
Lombardia	19	0.310	67%	0.646	53%
Marche	10	0.317	66%	0.707	72%
Molise	5	0.336	68%	0.740	100%
Piemonte	15	0.314	68%	0.673	58%
P.A. Bolzano	9	0.347	82%	0.772	85%
P.A. Trento	22	0.354	82%	0.795	92%
Puglia	11	0.335	76%	0.740	74%
Sardegna	8	0.334	78%	0.740	69%
Sicilia	10	0.340	74%	0.779	86%
Toscana	6	0.297	60%	0.661	58%
Umbria	10	0.337	76%	0.730	73%
Valle d'Aosta	9	0.304	62%	0.676	47%
Veneto	13	0.331	76%	0.672	59%
mean	13.70	0.32	70.72%	0.72	71.35%
std.dev.	6.35	0.02	0.08	0.05	0.15

In the Correlation matrix of relatedness indicators there is a very high correlation between ARI and RSI.

The interesting fact is that the indices measuring actual relatedness are not correlated with the ones measuring relatedness of the technological domains chosen by Italian regions. In principle, a region should choose technological domains:

ii) in which it is specialized (i.e. it has competence and know-how)

iii) that are closely related to each other (i.e. with high proximity).

Therefore, if regions had chosen with these criteria in mind the indices based on actual specialisation should be lower than the indices based on the chosen domains.
The difference between the two sets of indices is close to zero on average. This means that regions have clearly chosen domains in which they already have a high degree of specialisation rather than trying to maximize the relatedness between the domains.

8.3 Connectivity empirical results

A first empirical assessment of strategic cross-border and trans-regional cooperation among European regions is based on the similarity measure, i.e. the degree of resemblance between two regions.
The following table provides an extract of the similarity results for pairs of Italian regions relative to both of the two correlation measures used.

Table 10 – Similarity index

Region i	Region j	Pearson	Tetrachoric
Veneto	Emilia-Romagna	0.503	0.713
Veneto	Marche	0.437	0.635
Piemonte	Lombardia	0.436	0.639
Provincia Autonoma di Trento	Friuli-Venezia Giulia	0.424	0.627
Veneto	Friuli-Venezia Giulia	0.422	0.629
Provincia Autonoma di Bolzano	Provincia Autonoma di Trento	0.408	0.606
Lombardia	Emilia-Romagna	0.407	0.614
Provincia Autonoma di Bolzano	Veneto	0.404	0.604
Provincia Autonoma di Trento	Veneto	0.402	0.597
…	…	…	…
Molise	Lazio	-0.093	-0.173
Lombardia	Basilicata	-0.102	-0.173
Basilicata	Lazio	-0.107	-0.193
Liguria	Umbria	-0.132	-0.214
Valle d'Aosta	Campania	-0.151	-0.283

The empirical results suggest a strong **economic sector effect** on the resemblance. Regions with similar structures and importance of economic sectors on regional GDP show similar technological specialisation as measured by their patent production. For example, the pairs topping the list are among the more developed regions in Italy, all with an important share of manufacturing in total GDP.
However, these results only partly match the aims of S3 strategy for interaction between European regions in sharing and exchanging research and innovation. All the regions at the top of the list have a higher level of economic and technological development whereas all the less developed regions are at the bottom of the list with even negative values of

similarity. Molise, Basilicata and Campania have the lowest levels of similarity when coupled with more developed regions.

Compared to expectations based on the S3 framework, the results are counterintuitive. One aim was to promote connections between regions at different development levels with possible spillover effects from the more developed to those lagging behind. This should be possible within a Member State, but results suggest that connectivity in terms of similarity between regions at different levels of development is low and tends to be even lower between regions in different Member States.

A second connectivity indicator tries to measure if integrating specialisation between two regions would benefit both compared to their standalone capacity.

This approach is more in line with the S3 idea as connectivity is maximized when two regions reciprocally combine different specialisation and technological knowledge, increasing their overall innovation capacity.

In practical terms, the complementarity measure could be considered as a measure of relatedness calculated on pairs of regions. This is an attempt to measure how much the relatedness of region *k* alone would increase when combined with region *i*.

For this reason, connectivity in terms of complementarity is the average proximity between each pair of IPC classes in which the 'augmented' region is specialized.

Table below shows that in this case there is a strong **'innovation hub' effect** for each pair of regions. The most innovative region (Emilia-Romagna) strongly affects the overall complementarity index as it maximizes the index for many other regions.

Table 11 – Complementarity index

Region i	Region j	Index
Provincia Autonoma di Bolzano	Provincia Autonoma di Trento	0.821
Valle d'Aosta	Emilia-Romagna	0.807
Sardegna	Emilia-Romagna	0.807
Sicilia	Emilia-Romagna	0.807
Provincia Autonoma di Trento	Emilia-Romagna	0.806
Molise	Emilia-Romagna	0.805
Abruzzo	Provincia Autonoma di Trento	0.804
Basilicata	Emilia-Romagna	0.804
Calabria	Emilia-Romagna	0.802
…	…	…
Valle d'Aosta	Basilicata	0.557
Molise	Toscana	0.555
Valle d'Aosta	Liguria	0.549
Basilicata	Toscana	0.546

However, complementarity appears to be asymmetric. Even if the other regions have big incentives to collaborate with the hub, the opposite is not true.

Compared to the relatedness level of the region alone, the 'augmented relatedness' depends on the region. The following matrix of increments shows the increase/reduction between each pair of Italian regions.

As visible in the table, almost all the Italian regions have an advantage (increase) in collaboration with Emilia-Romagna compared with their standalone situation (vertical highlighted percentages) but there is no advantage for Emilia-Romagna in collaborating with any of the other regions (horizontal highlighted percentages).

In practical terms, the average proximity of Emilia Romagna decreases when combined with other regions, i.e. the relatedness within the specialisation domains of the region tends to decrease with the introduction of another region.

It is interesting to note that the greatest advantage in collaboration with Emilia-Romagna would be for Basilicata (+40.21%), Campania (+36.81%) and Valle d'Aosta (+38.35%). The small size of Valle d'Aosta probably affects the results more than the level of its economic development.

Table 12 – Matrix of complementarity increases and reductions

	Abruzzo	Basilicata	Bolzano	Calabria	Campania	Emilia-Romagna	Friuli-Venezia Giulia	Lazio	Liguria	Lombardia	Marche	Molise	Piemonte	Puglia	Sardegna	Sicilia	Toscana	Trento	Umbria	Valle d'Aosta	Veneto
Abruzzo		-1.98%	8.66%	0.29%	0.08%	12.43%	7.50%	-8.29%	-7.26%	-6.77%	-5.94%	-2.56%	-5.80%	7.33%	-0.66%	-5.57%	-20.70%	13.73%	-4.14%	-7.13%	1.42%
Basilicata	20.87%		34.55%	13.46%	0.56%	40.21%	33.45%	3.20%	0.53%	11.49%	13.14%	10.93%	14.77%	23.11%	30.23%	12.50%	-4.71%	35.20%	24.08%	-2.78%	28.88%
Bolzano	-2.75%	-2.34%		-2.44%	-14.92%	1.01%	-4.14%	-21.55%	-16.54%	-18.37%	-9.11%	-2.12%	-15.96%	-9.06%	-7.84%	-9.02%	-26.67%	3.93%	-9.87%	-5.72%	-5.70%
Calabria	2.52%	-5.94%	11.43%		-12.39%	15.98%	5.19%	-10.12%	-17.76%	-6.07%	-5.42%	-4.14%	-5.11%	-5.10%	-7.95%	-8.18%	-18.09%	13.00%	-3.16%	-9.30%	7.02%
Campania	24.08%	1.11%	17.86%	6.27%		36.81%	16.96%	4.44%	0.68%	15.90%	8.20%	0.59%	6.37%	12.64%	6.64%	6.55%	-0.92%	15.48%	17.54%	3.11%	22.77%
Emilia-Romagna	-1.12%	0.00%	-0.74%	-0.22%	-2.96%		-2.76%	-4.54%	-12.01%	-13.20%	-2.04%	0.10%	-11.35%	-1.35%	0.28%	0.28%	-12.56%	0.17%	-2.21%	0.30%	-10.84%
Friuli-Venezia Giulia	0.59%	1.25%	0.21%	-3.72%	-11.74%	3.45%		-12.55%	-13.73%	-15.41%	-8.76%	-0.63%	-12.48%	-8.81%	-3.22%	0.07%	-22.82%	3.98%	-9.44%	-9.37%	-3.94%
Lazio	10.98%	1.27%	6.06%	6.39%	1.92%	31.34%	13.09%		-1.13%	9.17%	1.41%	0.00%	0.93%	9.35%	11.72%	6.76%	16.41%	14.05%	9.14%	-4.20%	26.56%
Liguria	14.86%	0.96%	15.48%	-0.36%	0.57%	23.90%	14.18%	1.19%		13.41%	11.18%	2.08%	12.26%	1.82%	4.65%	10.83%	-0.91%	9.77%	8.00%	-3.92%	19.29%
Lombardia	2.40%	-0.70%	0.18%	0.93%	2.68%	8.40%	-0.70%	-0.91%	0.58%		3.25%	-0.70%	4.07%	0.56%	0.74%	-0.26%	6.01%	0.16%	0.00%	-0.70%	1.68%
Marche	4.58%	2.01%	12.91%	2.87%	-2.98%	23.84%	8.42%	-6.82%	-0.19%	4.52%		0.36%	3.28%	4.42%	6.43%	1.45%	-11.77%	9.24%	-2.52%	-2.40%	14.58%
Molise	5.55%	-2.56%	18.46%	1.58%	-12.12%	23.29%	15.04%	-10.48%	-10.71%	-2.07%	-2.23%		0.82%	12.24%	9.59%	11.13%	-14.91%	14.17%	7.68%	-7.18%	11.44%
Piemonte	1.21%	0.00%	0.88%	-0.27%	-7.82%	8.30%	0.50%	-10.38%	-2.61%	1.80%	-0.20%	0.00%		-2.45%	-0.99%	0.50%	-9.73%	2.29%	-1.96%	-0.62%	5.71%
Puglia	3.58%	-3.65%	-1.94%	-10.41%	-12.33%	8.25%	-5.95%	-12.79%	-20.66%	-11.65%	-9.37%	0.00%	-12.38%		0.47%	-7.28%	-22.56%	1.97%	-4.49%	-7.38%	0.58%
Sardegna	-5.79%	0.14%	-2.35%	-14.61%	-18.44%	8.13%	-1.92%	-12.45%	-19.87%	-13.03%	-9.23%	-4.06%	-12.62%	-1.27%		-7.53%	-23.69%	5.89%	-3.51%	-9.72%	-0.72%
Sicilia	-5.22%	-8.44%	2.02%	-9.85%	-13.75%	14.44%	7.34%	-11.45%	-10.18%	-8.86%	-8.43%	2.97%	-6.12%	-3.57%	-2.13%		-18.36%	6.02%	-5.70%	-10.63%	5.62%
Toscana	-0.16%	-2.70%	3.16%	0.89%	0.61%	25.18%	3.86%	21.14%	0.74%	21.52%	-0.08%	-1.09%	5.79%	1.03%	1.33%	2.42%		2.95%	2.64%	-0.15%	22.85%
Trento	-0.02%	-3.61%	2.09%	-2.81%	-18.12%	0.14%	-2.30%	-17.14%	-22.07%	-19.83%	-13.62%	-7.34%	-16.30%	-7.11%	-1.82%	-7.13%	-28.12%		-9.13%	-14.33%	-7.14%
Umbria	-6.63%	-1.99%	-1.92%	-7.73%	-7.67%	8.30%	-5.72%	-12.15%	-15.06%	-11.32%	-14.60%	-3.17%	-11.12%	-3.60%	-0.89%	-8.48%	-20.60%	0.68%		-8.43%	-2.10%
Valle d'Aosta	12.67%	-4.35%	27.79%	7.64%	0.89%	38.35%	17.51%	-3.95%	-5.88%	9.68%	6.50%	3.96%	12.22%	16.44%	15.50%	8.03%	-3.79%	18.22%	14.05%		26.79%
Veneto	-2.96%	0.00%	0.81%	0.18%	-5.25%	-2.99%	-1.76%	0.08%	-7.84%	-11.42%	-1.39%	-1.56%	-5.85%	-0.27%	0.18%	0.70%	-6.64%	1.07%	-3.83%	0.00%	

8.3.1 Connectivity and research projects

A second approach to measuring connectivity and validating the indices developed with patent data is based on a broader analysis of NUTS2 research activity. All the European research projects under FP7 available in the CORDIS database are used to assess the correlation of research collaboration of different regional actors with the connectivity measures developed in the previous section.

Analysis methods are comparable to those adopted for patents as both could be considered as outputs of the innovation process. An index of collaboration between each pair of all European NUTS2 is calculated as the number of projects in which at least one participant was from region *r* and one participant from region *s*. The $K \times K$ symmetric matrix of regions with the main diagonal set to 0 shows the total number of projects between any two regions as reported in CORDIS. Each cell (r,s) of this matrix represents the number of projects between organizations in regions *r* and *s*. For example, the maximum number of projects in collaboration is between Upper Bavaria (DE21) and Île de France (FR10) with 1312 co-occurrences. It is interesting to note that in this case the capital cities of these regions, Munich and Paris, probably affect the indicator even though the analysis is carried out at regional level.

The **collaboration index** could be written as:

$$C_{r,s} = \frac{\sum_{i=1}^{N} I(i \in p_{r,s})}{\sqrt{\sum_{i=1}^{N} I(i \in p_r) * \sum_{i=1}^{N} I(i \in p_s)}}$$

where $C_{r,s}$ is the index of collaboration between region *r* and region *s* and $p_{r,s}$ are the projects which involve both regions. The index is normalized by the square root of the product between the total number of projects in region *r* and the total number of projects in region *s*. For example, there are 19 projects with organizations belonging to both regions AT11 and AT12 and there are 2243 projects with organizations in AT11 and 1904 in AT12 then the index of collaboration between AT11 and AT12 is 19/√(2243*1904).

The collaboration index $C_{r,s}$ is used as a dependent variable in several regression models with any pair of European regions as observations (more than 30 000).

Table 13 – Summary statistics of collaboration index

Min	1st quartile	Median	Mean	3rd quartile	Max
0	0	0.00181	0.00265	0.00374	0.0635

Validating the patents measures uses the similarity and connectivity indices as covariates of the collaboration index with additional control variables for the relationship[44] included in three different models. The base model only considers the two indices of connectivity, a second model incorporates geographical controls as suggested in the literature and finally a third model includes also economic development controls.

Specifically, the variables in the models are:

- ✓ *d_sameMS*: dummy variable for both regions being in the same Member State;
- ✓ *d_sameNUTS1*: dummy variable indicating if the regions are in the same NUTS1;
- ✓ *log(distance)*: logarithm of the haversine distance between the regions' centroids;
- ✓ *sim_pers*: similarity index (Pearson);
- ✓ *compl*: complementarity index;
- ✓ *inc_dec_abs*: absolute value of the relatedness increase/decrease;
- ✓ *diff_pop*: absolute value of the difference between populations of each pair in 2016;
- ✓ *diff_gdp*: absolute value of the difference between GDPs of each pair in 2016;
- ✓ *diff_pat*: absolute value of the difference between the number of patents in 2012.

There was a preliminary check on collinearity between covariates to exclude redundant variables affecting the coefficient results.

The OLS models presented in the following page reveal a positive relationship between the collaboration index and both geographic dummy variables indicating when the pair of regions is in the same Member State and the same NUTS1. It is interesting to note that the parameter is greater for the same NUTS1 than for the same Member State highlighting the importance of proximity between regions sharing projects.

In addition, the distance variable, which is the logarithm of the haversine distance between region centroids, confirms the importance of geographical proximity. The collaboration index tends to decrease when the distance increases.

The positive parameters of log-distance squared suggests that this relationship is non-linear and less than proportional, i.e. there is a threshold of distance beyond which the collaboration index decrease is negligible[45].

[44] Population, GDP and patent data are from the Eurostat web service https://ec.europa.eu/eurostat/web/regions/data/database, in particular the endpoints demo_r_d2jan, nama_10r_2gdp, pat_ep_rtot

[45] y = -0.00293*log(x)+0.00009*log(x)^2

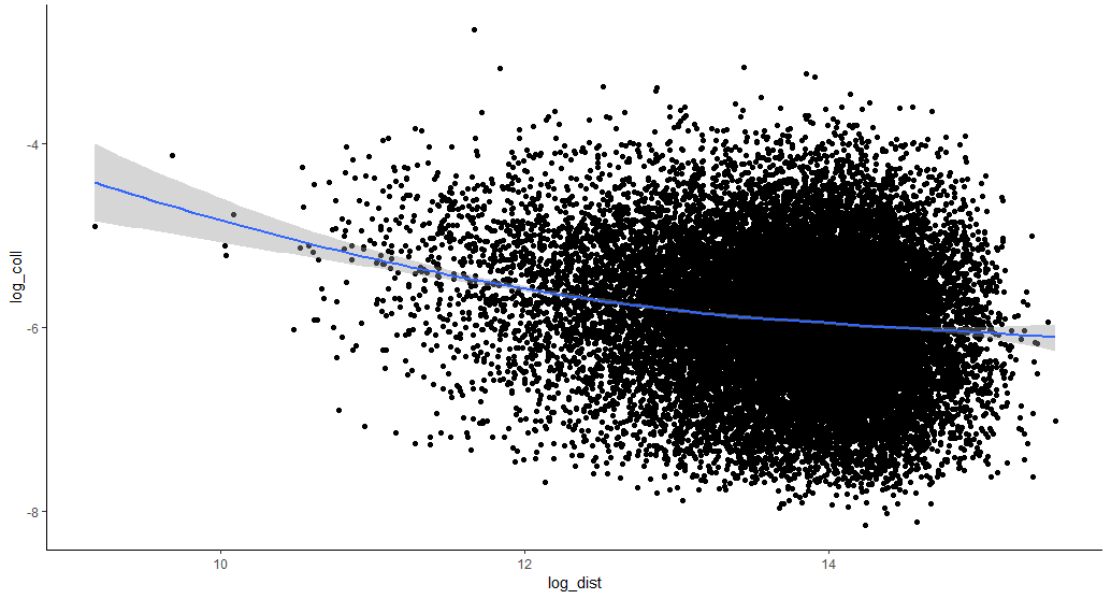
Table 14 – Regression results: index of collaboration as dependent

	Model 1	Model 2	Model 3
(Intercept)	0.00737 ***	0.03029 ***	0.03832 ***
	(0.00011)	(0.00452)	(0.00429)
sim_pers	**0.00196 *****	**0.00119 *****	**0.00159 *****
	(0.00031)	**(0.00031)**	**(0.00029)**
compl	**-0.00782 *****	**-0.00735 *****	**-0.00561 *****
	(0.00019)	**(0.00019)**	**(0.00019)**
d_sameMS		0.00074 ***	0.00073 ***
		(0.00008)	(0.00008)
d_sameNUTS1		0.00146 ***	0.00152 ***
		(0.00019)	(0.00018)
log(distance)		-0.00293 ***	-0.00440 ***
		(0.00066)	(0.00063)
log(distance)^2		0.00009 ***	0.00014 ***
		(0.00002)	(0.00002)
abs(inc_dec_abs)		0.00048 **	
		(0.00017)	
diff_pop			0.00000 ***
			(0.00000)
diff_gdp			0.00000 ***
			(0.00000)
diff_pat			0.00000 ***
			(0.00000)
R^2	0.05403	0.08799	0.17239
Adj. R^2	0.05397	0.08778	0.17215
Num. obs.	30628	30628	30628
RMSE	0.00307	0.00301	0.00287

*** p < 0.001, ** p < 0.01, * p < 0.05

This negative and significant relationship between collaboration and distance is also shown in the following figure and is consistent with the literature (Cecere & Corrocher, 2015).

Figure 41 – The index of collaboration and the spatial distance between regions (in log)



The two connectivity indices based on patents show opposite results. While collaboration increases when the similarity index increases, there is a negative relationship with complementarity, which means there is less integration of specialisation in technological domains.

This suggests that collaborations between different regions are (usually) based on similarities whereas complementarities within the same region (relatedness) are unlikely to be the basis for collaboration with other regions.

Distance matters when connecting different technological domains but is less relevant for collaboration between similar regions. When one region is trying to benefit from variety, spatial proximity matters.

The empirical results provide an interesting suggestion, that connectivity between regions based on complementarity is much more difficult to exploit than connectivity based on similarities. Therefore, contrary to the aims of S3 to promote collaboration between regions at different levels of development, collaboration seems to be more likely only with similarities in the structure of innovation.

The other control variables for the population, GDP and patents, are proxies of regional size and are likely to affect the collaboration index. It is interesting to note that large differences in population, as well as economic and technological development positively affects collaboration, suggesting frequent interaction between large and small regions.

Finally, it is interesting to note the substantial increase of adjusted $R^{\wedge 2}$ resulting when including the control variables.

There are comparable results for parameters and significance when using the logarithm of the collaboration index as shown in the following table.

Table 15 – Regression results: log(index of collaboration) as dependent

	Model 1	Model 2	Model 3
(Intercept)	-5.12479 ***	-1.14761	0.45794
	(0.03159)	(1.24999)	(1.21396)
sim_pers	**0.54553 *****	**0.37537 *****	**0.43260 *****
	(0.08871)	**(0.08843)**	**(0.08607)**
compl	**-1.31537 *****	**-1.26055 *****	**-1.07514 *****
	(0.05266)	**(0.05482)**	**(0.05542)**
d_sameMS		0.18208 ***	0.18339 ***
		(0.02235)	(0.02176)
d_sameNUTS1		0.29329 ***	0.30080 ***
		(0.05100)	(0.04961)
log(distance)		-0.49633 **	-0.77171 ***
		(0.18462)	(0.17967)
log(distance)^2		0.01471 *	0.02483 ***
		(0.00683)	(0.00664)
abs(inc_dec_abs)		0.06405	
		(0.05013)	
diff_pop			0.00000 ***
			(0.00000)
diff_gdp			0.00000 ***
			(0.00000)
diff_pat			0.00008 ***
			(0.00002)
R^2	0.02753	0.05637	0.10600
Adj. R^2	0.02744	0.05608	0.10564
Num. obs.	22330	22330	22330
RMSE	0.74786	0.73676	0.71716

*** p < 0.001, ** p < 0.01, * p < 0.05

9 Final considerations

This research is based on extensive available open data. The vast unused potential and unexploited informative power of this data is especially important given that budgetary constraints require European regional development policies to demonstrate their added value and importance with evidence, both ex ante for decision making and ex post for assessment of the outcome.

Contrary to literature, open data is used to develop methodologies and tools for decision making. In particular, this research emphasizes the importance of staying ahead. Regular monitoring helps keep the evolution of spending and results under control as ex post analysis cannot fix critical situations. The research also shows how augmenting a data structure relative to specific issues supports decision making.

In so doing the research addresses the first hypothesis:

> ***H1****: Open data platforms can be considered useful for policy making and not just as data tombs set up only to satisfy governmental digital agenda requirements.*

As opposed to literature and state-of-the-art tools, the analytics and visualization in ESIFy keep the user experience simple but widen access to cover the full set of data available.

The dashboard design allows multiple visualizations in each panel and displays several variables at the same time for each of these views.

This implies increased information in terms of:

- ✓ figures;
- ✓ measures and dimensions per figure;
- ✓ units of observations (e.g. EU, Member State, regions) per figure.

Most importantly, the major innovation is that the design is based on **benchmarking**, with comparisons of KPI relative performance in different aggregates (over time and space).

Simple visualizations make explicit what is already in the data, which is transformed into knowledge for evidence-based decision making.

The views do not require robust analytical skills to easily see:

- things that are known and have to be demonstrated, e.g. delays in program implementation. Usually, MAs can see implementation progress and its comparison to targets. However, this 'self-oriented' information is marginally useful as both the implementation and the target are region-specific. The benchmarking approach allows comparisons of this position against reference MSs and other OPs. This is

especially true with the level playing field effect of the Partnership Agreement and country-specific conditions applied to all regions.

- things that are not known and can be discovered, e.g. low targets. This could reflect at least one of the following reasons:

  1) given the performance framework in the regulation, underestimation has been a way for program authorities to reduce the probability of being penalized by a review of their performance;

  2) low capacity of MAs (and lack of methodologies) to measure reasonable final targets;

  3) low capacity and lack of methodology to argue that these targets were not set appropriately in the ex-ante phase.

In this sense, the data should be understood as **smart data**, i.e. available and used in an intelligent way to show trends, regularities and irregularities.

The chapters go into more detail but other examples of insight on ESIF data are:

- highly mature project selection in some European MS which positively impacts disbursement;
- uneven priority axes progress within the same Operational Programmes.

The power of data visualization reveals trends when combining **benchmarking** with the **multidimensional** framework in the dashboard. Simply exposing large amounts of data for digital agenda requirements does not produce comparable informative effects.

The research also addresses a second hypothesis:

> ***H2:*** *In allocating thematic objective 1 investments, regions have developed their S3s according to embeddedness, relatedness and connectivity.*

In general, regions have chosen about a third of the domains in which they show relative specialisation (i.e. RCA>0). In other words, there is a large share of 'owned but not chosen' with substantial differences around the mean.

However, as regards the specific coherence exercise, it is interesting to measure the chosen domains where the region has a relative advantage (as measured by the RCA index).

On average, the degree of 'coherence' is slightly less than 50%.

Comparing Italian regions, contrary to expectations many of them have a narrow span of specialisation but low coherence, whereas a few other, despite a larger span of specialisation, have a level of coherence above the mean.

When coherence index includes the potential technological domain (positive patenting trend in the recent years) only a few regions show significant increase in the coherence indicator, i.e. policy orientation was mainly toward current rather than potential technological domains.

Some regions have few domains in which they show absolute strength but have nonetheless chosen a large span of specialisation, resulting in a low level of the coherence indicator.

All relatedness indices are negatively correlated with the number of chosen domains. Intuitively the larger the number of domains the lower their degree of relatedness, i.e. the higher the number of knowledge fields, the higher their 'distance'.

In the view of coherence results, it appears that regions have clearly chosen domains in which they already have a high degree of (current) specialisation rather than trying to maximize the relatedness between the domains.

An attempt to measure connectivity among regions was carried out developing the similarity and complementarity indexes.

The empirical results suggest a strong 'economic sector' effect on the similarity.

Regions with similar structures and importance of economic sectors on regional GDP show similar technological specialisation as measured by their patent production.

On the contrary, using complementarity implies a strong 'innovation hub' effect for each pair of regions: the most innovative region (Emilia-Romagna) strongly affects the overall complementarity index as it maximizes the index for many other regions.

Furthermore, complementarity appears to be asymmetric. Even if the other regions have big incentives to collaborate with the hub, the opposite is not true.

When validating the connectivity indexes against CORDIS research project data, a positive relationship between the collaboration index and geographic proximity.

It is interesting to note that the regression coefficient is greater for the same NUTS1 than for the same Member State highlighting the importance of proximity between regions sharing projects.

Furthermore, the distance variable confirms the importance of geographical proximity: the connectivity index tends to decrease when the geographical distance increases.

Additionally, the positive coefficient of log-distance squared suggests that this relationship is non-linear and less than proportional, i.e. there is a threshold of distance beyond which the collaboration index decrease is negligible.

The two connectivity indices based on patents show opposite results: while collaboration increases when the similarity index increases, there is a negative relationship with complementarity, which means there is less integration of specialisation in technological

domains. This empirical result provide an interesting suggestion: connectivity between regions based on complementarity is much more difficult to exploit than connectivity based on similarities.

Therefore, contrary to the aims of S3 to promote collaboration between regions at different levels of development, collaboration seems to be more likely only with similarities in the structure of innovation.

Despite the empirical results could not be extended Europe wide, the Italian testbed regions seem to only partially follow S3 guide recommendations, in particular:

- **Embeddedness:** indices show a generally acceptable level despite extensive differences due to regional sizes and levels of development. However, the entrepreneurial discovery approach does not fully result in a high level of coherence between existing or promising specializations and choices. Furthermore, the indices show many 'owned but not chosen' technological domains. In addition, adding promising domains significantly increased coherence in only a few regions.
- **Relatedness:** the low level of the index suggests that regions chose domains where they already have a high degree of specialization (coherence) rather than trying to maximize relatedness between domains. Furthermore, some regions did not significantly improve relatedness compared to a completely random choice.
- **Connectivity:** the S3 approach required regions to identify potential relations with other EU regions, possibly with different levels of development. However, the empirical evidence shows:
  - ✓ collaboration between different regions is (usually) based on similarities;
  - ✓ complementarities within a region (relatedness) are unlikely to be the basis for collaboration with other regions.

The results suggest that regions usually adopt a prudent approach in selecting technological domains, due in part to the absence of a clear methodology and indications. Connectivity between regions based on complementarity is much more difficult to exploit than connectivity based on similarities. Therefore, contrary to S3 aims of promoting collaboration between regions at different levels of development, collaboration seems more likely only when regions have similar innovation structures.

However, according to the European Commission draft regulation proposal of May 2018[46] for the next 2021-2027 programming period, research and innovation (TO1) and support

[46] https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52018PC0375&from=EN
https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52018PC0372&from=EN

to SMEs (TO3) will be merged in the policy objective 1 *'A smarter Europe promoting innovative and smart economic transformation'* with strong emphasis on **international collaboration**.

As described in Annex IV, the *enabling condition* S3 will be supported by criteria in the following table integrated with the proposed common indicators (see also *Development of a system of common indicators for European Regional Development Fund and Cohesion Fund interventions after 2020*[47].)

[47] http://ec.europa.eu/regional_policy/en/information/publications/studies/2018/development-of-a-system-of-common-indicators-for-european-regional-development-fund-and-cohesion-fund-interventions-after-2020-part-i-thematic-objective-1-3-4-5-6

Table 16 – Overview of the new regulation for research and innovation

CPR Annex IV[48]				ERDF Annex I[49]	
Policy objective	**Specific objective**	**Enabling condition**	**Fulfillment criteria for the enabling condition**	**Common output indicators**	**Common result indicators**
1 A smarter Europe promoting innovative and smart economic transformation	(i) enhancing research and innovation capacities and the uptake of advanced technologies; (ii) reaping the benefits of digitisation for citizens, companies and governments; (iii) enhancing growth and competitiveness of SMEs; (iv) developing skills for smart specialisation, industrial transition and entrepreneurship.	Good governance of national or regional **smart specialisation strategy**	Smart specialisation strategy shall be supported by: 1. Up-to-date analysis of bottlenecks for innovation diffusion, including digitisation 2. Existence of competent regional / national institution or body, responsible for the management of the smart specialisation strategy 3. Monitoring and evaluation tools to measure performance towards the objectives of the strategy 4. Effective functioning of entrepreneurial discovery process 5. Actions necessary to improve national or regional research and innovation systems 6. Actions to manage industrial transition 7. Measures for international collaboration	RCO 01 - Enterprises supported (large and MSMEs) RCO 02 - Enterprises supported by grants RCO 03 - Enterprises supported by financial instruments RCO 04 - Enterprises with non-financial support RCO 05 - Start-ups supported RCO 06 - Researchers working in supported research facilities RCO 07 - Research institutions participating in joint research projects RCO 08 - Nominal value of research and innovation equipment RCO 10 - Enterprises cooperating with research institutions **RCO 96 – Interregional investments in EU projects**	RCR 01 - Jobs created in supported entities RCR 02 - Private investments matching public support (of which: grants, financial instruments) RCR 03 – SMEs introducing product or process innovation RCR 04 - SMEs introducing marketing or organisational innovation RCR 05 - SMEs innovating in-house **RCR 06 - Patent applications submitted to European Patent Office** RCR 07 - Trademark and design applications RCR 08 - Public-private co-publications

[48] https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52018PC0375&from=EN
[49] https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52018PC0372&from=EN

Beyond the measure of collaboration and in a wider perspective, the regulation explicitly requires monitoring and evaluation tools to measure performance towards strategy objectives (point 3). In this sense, the following table tentatively matches the current principles (and measures developed in this research) with the future regulation.

Table 17 – Current and future S3

S3 principles	S3 fulfilment criteria
Embeddedness	4. Effective functioning of entrepreneurial discovery process
Relatedness	5. Actions necessary to improve national or regional research and innovation systems
Connectivity	7. Measures for international collaboration

This research is based on patent data and result indicators include applications. However, using patents to measure innovation can only partially reflect innovation potential, especially in Italian manufacturing regions.
Moreover, the distribution of patents is highly concentrated as not all industries and firms rely on patents when producing and applying new knowledge. In addition, a few large companies own most of the patents. This is a problem especially for lagging regions, with small firms in low-tech industries not using patents to protect their knowledge at all.

The research presented in this document could not solve two aspects, which leaves space for further research and methodology validation.
Firstly, applying the methodologies to EU Member States other than Italy is limited by the development of S3 documents in national languages.
Secondly, and of foremost importance, is the availability and adoption of project level data. Although project level data is unlikely to be in a structured format, it is currently the only available source of micro level information.
Micro data at MA level on financed projects aggregated for AIR submission, is undoubtedly the most fundamental asset for monitoring and decision making as it gives access to crucial information at ground level. Data on the types of beneficiaries, economic sectors, amount of projects in EUR, duration of implementation and geographic localization could exponentially increase the informative capacity of data and advise policy.
These data could lead to an easy expansion of the information base joining, for example, patent data by geographical match.
This would allow the measures proposed in this research to be applied to the deepest level available and granular assessment of the degree of S3 principles. This is particularly true

considering the nature of selection criteria in the calls. Broad selection criteria enable more applications, including from projects not necessarily oriented to policy objectives. Stricter selection criteria allow better adherence to policy strategy but a lower participation rate.

However, MAs could be reticent to disseminate project data due to the additional administrative burden related to the larger information size and longer monitoring times. As usual, EU incentives for the regional counterpart would ease the adoption of project level information reporting. Examples of incentives range from reducing audit procedures on resource management to increasing planned resources with proportional premia.

Smart data are awaiting to be unveiled.

10 References

Bibliography

Acs, Z. J., Anselin, L., & Varga, A. (2002). Patents and Innovation Counts as Measures of Regional Production of New Knowledge. *Research Policy*, *31*(7), 1069–1085.

Anjali, M. K., & P, B. A. (2014). Ambiguities in Natural Language Processing. *International Journal of Innovative Research in Computer and Communication Engineering*, 392–394.

Asheim, B. T., Boschma, R., & Cooke, P. (2011). Constructing Regional Advantage: Platform Policies Based on Related Variety and Differentiated Knowledge Bases. *Regional Studies*, *45*(7), 893–904. https://doi.org/10.1080/00343404.2010.543126

Asheim, B. T., & Grillitsch, M. (2015). Smart specialisation: Sources for new path development in a peripheral manufacturing region. *Papers in Innovation Studies No. 2015/11, CIRCEL; Lund Univeristy*, 1–27.

Balland, P. A., Boschma, R., Crespo, J., & Rigby, D. L. (2018). Smart specialization policy in the European Union: relatedness, knowledge complexity and regional diversification. *Regional Studies*, *0*(0), 1–17. https://doi.org/10.1080/00343404.2018.1437900

Balland, P. A., & Rigby, D. (2017). The Geography of Complex Knowledge. *Economic Geography*, *93*(1), 1–23. https://doi.org/10.1080/00130095.2016.1205947

Bateman, S., Mandryk, R. L., Gutwin, C., Genest, A., Mcdine, D., & Brooks, C. (2010). Useful Junk ? The Effects of Visual Embellishment on Comprehension and Memorability of Charts. *Human Factors*, 2573–2582. https://doi.org/10.1145/1753326.1753716

Berends, J., Carrara, W., Engbers, W., & Vollers, H. (2017). Re-Using Open Data, 106. Retrieved from https://www.europeandataportal.eu/sites/default/files/re-using_open_data.pdf

Berners-Lee, T. (2013). Linked data.

Birkland, T. A. (2014). *An introduction to the policy process: Theories, concepts and models of public policy making*.

Borkin, M. A., Vo, A. A., Bylinskii, Z., Isoa, P., Sunkavalli, S., Oliva, A., & Phister, H. (2013). What makes a data visualization memorable ? *IEEE Transactions on Visualization & Computer Graphics*, *19*(12), 2306–2315. https://doi.org/10.1109/TVCG.2013.234

Boschma, R., Eriksson, R., & Lindgren, U. (2009). How does labour mobility affect the performance of plants? The importance of relatedness and geographical proximity. *Journal of Economic Geography*, *9*(2), 169–190. https://doi.org/10.1093/jeg/lbn041

Boschma, R., & Frenken, K. (2011a). Technolgical relatedness, related variety and economic geography. In P. Cooke, B. T. Asheim, R. Boschma, R. Martin, D. Schwartz, & F. Todtling (Eds.), *Handbook of Regional Innovation and Growth* (Vol. 11, pp. 187–197). Cheltenam (UK): Edward Elgar. https://doi.org/10.1093/jeg/lbq053

Boschma, R., & Frenken, K. (2011b). Technological relatedness and regional branching. In H. Bathelt, M. Feldman, & D. F. Kogler (Eds.), *Beyond Territory: Dynamic Geographies of Knowledge Creation, Diffusion and Innovation* (pp. 64–81). London: Routledge.

Boschma, R., & Frenken, K. (2011c). The emerging empirics of evolutionary economic geography. *Journal of Economic Geography*, *11*(2), 295–307. https://doi.org/10.1093/jeg/lbq053

Boschma, R., & Gianelle, C. (2013). Regional branching and smart specialization policy.

JRC Technical Reports, (06/2104). https://doi.org/http://dx.doi.org/10.2791/65062

Boschma, R., & Gianelle, C. (2014). Regional Branching and Smart Specialisation Policy. *S3 Policy Brief Series*, (06). https://doi.org/10.2791/65062

Boschma, R., & Iammarino, S. (2009). Related Variety, Trade Linkages, and Regional Growth in Italy. *Economic Geography*, *85*(3), 289–311. https://doi.org/10.1111/j.1944-8287.2009.01034.x

Boschma, R., Minondo, A., & Navarro, M. (2010). *Related variety and regional growth in Spain*. *Papers in Evolutionary Economic Geography*. Utrecht: Urban & Regional Research Centre Utrecht.

C.Hood. (2007). Intellectual Obsolescence and Intellectual Makeovers: Reflections on the Tools of Government after Two Decades. *Governance*.

Cairo, A. (2015). New Challenges for Data Design, 103–116. https://doi.org/10.1007/978-1-4471-6596-5

Camagni, R., & Capello, R. (2013). Regional innovation patterns and the eu regional policy reform: Toward smart innovation policies. *Growth and Change*, *44*(2), 355–389.

Caragliu, A., & Del Bo, C. (2018). Much Ado About Something? An Appraisal of the Relationship Between Smart City and Smart Specialisation Policies. *Tijdschrift Voor Economische En Sociale Geografie*, *109*(1), 129–143. https://doi.org/10.1111/tesg.12272

Cecere, G., & Corrocher, N. (2015). The Intensity of Interregional Cooperation in Information and Communication Technology Projects: An Empirical Analysis of the Framework Programme. *Regional Studies*, *49*(2), 204–218. https://doi.org/10.1080/00343404.2012.759651

Dalum, B., Laursen, K., & Villumsen, G. (1998). Structural Change in OECD Export Specialisation Patterns: de-specialisation and "stickiness." *International Review of Applied Economics*, *12*(3), 423–443.

Dawes, S. S. (2010). Stewardship and usefulness: Policy principles for information-based transparency. *Government Information Quarterly*, *27*(4), 377–383. https://doi.org/10.1016/j.giq.2010.07.001

Dawes, S. S., & Helbig, N. (2010). Information Strategies for Open Government: Challenges and Prospects for Deriving Public Value from Government Transparency BT - Electronic Government: 9th IFIP WG 8.5 International Conference, EGOV 2010, Lausanne, Switzerland, August 29 - September 2, 2, 50–60. https://doi.org/10.1007/978-3-642-14799-9_5

Duranton, G., & Puga, D. (2001). Nursery Cities: Urban Diversity, Process Innovation, and the Life Cycle of Products. *American Economic Review*, *91*(5), 1454–1477.

Erl, T., Khattak, W., & Buhler, P. (2016). *Big Data Fundamentals*. *Software Quality Professional* (Vol. 18).

ESPON. (2018). Potentials of big data for integrated territorial policy development in the European growth corridors (Big Data & EGC).

European Commission. (2017). Enter the Data Economy: EU Policies for a Thriving Data Ecosystem. *EPSC Strategic Notes*, (21), 16. https://doi.org/10.2872/5437

Fall, C. J., Benzineb, K., & Guyot, J. (2003). Computer-Assisted Categorization of Patent Documents in the International Patent Classification, *2003*(October), 1–14.

Fawcett, T. (2014). Data Science and Its Relationship to Big Data and Data-Driven Decision Making Data Science and its relationship to Big Data and data-driven decision making,

(March 2013). https://doi.org/10.1089/big.2013.1508

Few, S. (2006). *Information Dashboard Design The Effective Visual Communication of Data*.

Foray, D. (2015). *Smart Specialisation. Opportunities and Challenges for Regional Innovation Policy*. London: Routledge.

Foray, D. (2016). On the policy space of smart specialization strategies. *European Planning Studies*, *24*(8), 1428–1437. https://doi.org/10.1080/09654313.2016.1176126

Foray, D., David, P. A., & Hall, B. H. (2009). *Smart Specialisation – The Concept*. Brussels.

Foray, D., David, P. A., & Hall, B. H. (2011). *Smart specialisation - From academic idea to political instrument, the surprising career of a concept and the difficulties involved in its implementation*. *MTEI-Working_paper-2011-001*.

Foray, D., Goddard, J., Beldarrain, X. G., Landabaso, M., McCann, P., Morgan, K., … Mulatero, F. (2012). *Guide to Research and Innovation Strategies for Smart Specialisations (RIS 3)*. Brussels.

Foray, D., & Goenega, X. (2013). *The Goals of Smart Specialisation* (S3 Policy Brief Series). https://doi.org/10.2791/20158

Frenken, K., Van Oort, F., & Verburg, T. (2007a). Related variety, unrelated variety and regional economic growth. *Regional Studies*, *41*(5), 685–697. https://doi.org/10.1080/00343400601120296

Frenken, K., Van Oort, F., & Verburg, T. (2007b). Related Variety, Unrelated Variety and Regional Economic Growth. *Regional Studies*, *41*(5), 685–697.

Gascó-Hernández, M., Martin, E. G., Reggi, L., Pyo, S., & Luna-Reyes, L. F. (2018). Promoting the use of open government data: Cases of training and engagement. *Government Information Quarterly*, *35*(2), 233–242. https://doi.org/10.1016/j.giq.2018.01.003

Giest, S. (2017). Big data for policymaking : fad or fasttrack ? *Policy Sciences*. https://doi.org/10.1007/s11077-017-9293-1

Golfarelli, M., Maio, D., & Rizzi, S. (1998). THE DIMENSIONAL FACT MODEL : A CONCEPTUAL MODEL FOR DATA WAREHOUSES 1. *International Journal of Cooperative Information Systems*.

Grant, C. (2016). Supporting a Passion for New Ideas through Open APIs. *Information Services and Use*, *36*(1–2), 65–72. https://doi.org/10.3233/ISU-160798

Graves, Alvaro; Hendler, J. (2014). A study on the use of visualizations for Open Government Data. *Information Polity*.

Grillitsch, M., Asheim, B., & Trippl, M. (2018). Unrelated knowledge combinations: the unexplored potential for regional industrial path development. *Cambridge Journal of Regions, Economy and Society*, *11*(2), 257–274. https://doi.org/10.1093/cjres/rsy012

Helbig, N., Cresswell, A.M., Burke, G.B. and Luna-Reyes, L. (2012). The Dynamics of Opening Government Data: A White Paper. Centre for Technology in Government, State University of New York, Albany.

Hemerly, J. (2013). Public Policy Considerations for Data-Driven Innovation. *Computer*. https://doi.org/10.1109/MC.2013.186

Hidalgo, C. A., Klinger, B., Barabasi, A.-L., & Hausmann, R. (2007). The Product Space Conditions the Development of Nations. *Science*, *317*(5837), 482–487.

https://doi.org/10.1126/science.1144581

Hullman, J., Adar, E., & Shah, P. (2011). The impact of social information on visual judgments. *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems - CHI '11*, 1461. https://doi.org/10.1145/1978942.1979157

Hullman, J., Drucker, S., Henry Riche, N., Lee, B., Fisher, D., & Adar, E. (2013). A deeper understanding of sequence in narrative visualization. *IEEE Transactions on Visualization and Computer Graphics*, *19*(12), 2406–2415. https://doi.org/10.1109/TVCG.2013.119

Iacobucci, D. (2014). Designing and Implementing a Smart Specialisation Strategy at Regional Level : Some Open Questions. *Scienze Regionali, Italian Journal of Regional Science*, *13*(1), 107–126.

Iacobucci, D., & Guzzini, E. (2016a). Relatedness and connectivity in technological domains: missing links in S3 design and implementation. *European Planning Studies*, *24*(8), 1511–1526. https://doi.org/10.1080/09654313.2016.1170108

Iacobucci, D., & Guzzini, E. (2016b). Relatedness and connectivity in technological domains: the "missing links" in S3 design and implementation. *European Planning Studies*, *24*(8), 1511–1526. https://doi.org/10.1080/09654313.2016.1170108

Jacobs, J. (1969). *The economy of cities*. New York: Random House.

Jaffe, A. (1986). Technological opportunity and spillover of R&D. *Amercican Economic Review*, *76*, 984–1001.

Janssen, M. A. Z. Y. C. (2012). Benefits, Adoption Barriers and Myths of Open Data and Open Government. *Information System Management*.

Klievink, B., & Cunningham, S. (2017). Big data in the public sector : Uncertainties and readiness, 267–283. https://doi.org/10.1007/s10796-016-9686-2

Lambooy, J. G., & Boschma, R. (2001). Evolutionary economics and regional policy. *Annals of Regional Science*, *35*(1), 113–132.

Martin, Erika G., PhD, MPH; Begany, Grace M., M. (2018). Transforming Government Health Data Into All-Star Open Data: Benchmarking Data Quality.

McCann, P., & Ortega-Argilés, R. (2013). Some practical elements associated with the design of an integrated and territorial place-based approach to EU cohesion policy. *Geography, Institutions and Regional Economic Performance*.

McCann, P., & Ortega-Argilés, R. (2015). Smart Specialization, Regional Growth and Applications to European Union Cohesion Policy. *Regional Studies*, *49*(8), 1291–1302.

McCool, D. (1995). *Public policy theories, models, and concepts: An anthology*. Pearson College Div,.

McKinsey. (2011). Big data : The next frontier for innovation , competition , and productivity, (May).

Merino Huerta, Mauricio, et al. (2010). *Problemas, decisiones y soluciones. Enfoques de política pública*.

Miller, K. W., & St, M. (2013). Big Data : New opportunities and new challenges. *Computer*, 22–24.

Neffke, F., Henning, M., & Boschma, R. (2011). How do regions diversify over time? Industry relatedness and the development of new growth paths in regions. *Economic Geography*, *87*(3), 237–265.

Nooteboom, B., Van Haverbeke, W., Duysters, G., Gilsing, V., & van den Oord, A. (2007).

Optimal cognitive distance and absorptive capacity. *Research Policy*, *36*(7), 1016–1034.

Noveck, B. S. (2012). Wiki Government: How Technology Can Make Government Better, Democracy Stronger, and Citizens More Powerful. Washington: Brookings Institution Press. 2009. *MedieKultur: Journal of Media and Communication Research*, *28*(52), 4. https://doi.org/10.7146/mediekultur.v28i52.5731

OECD. (2018). Open Government Data Report: Enhancing Policy Maturity for Sustainable Impact. *OECD Digital Government Studies, OECD Publishing*. Retrieved from https://doi.org/10.1787/9789264305847-en

Pickle, L. W., & Monmonier, M. (1997). How to Lie with Maps. *The American Statistician*, *51*(2), 206. https://doi.org/10.2307/2685420

Ponds, R., Van Oort, F., & Frenken, K. (2007). The geographical and institutional proximity of research collaboration. *Papers in Regional Science*, *86*(3), 423–443. https://doi.org/10.1111/j.1435-5957.2007.00126.x

Radu, G. Cecconi, C. (2018). Open Data Maturity in Europe.

Ramon Gil-Garcia, J. (2017). *Policy Analytics: Definitions, Components, Methods, and Illustrative Examples*.

Sano, K. M. and D. (1995). *Designing Visual Interfaces*.

Schintler, L. A. (2014). Big Data for Policy Analysis: The Good, The Bad, and The Ugly, *31*(4), 343–348.

Shneiderman, B. (1996). The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. *The Craft of Information Visualization*, 364–371. https://doi.org/10.1016/B978-155860915-0/50046-9

Tufte, E. R. (2001). *The Visual Display of Quantitative Information - 2nd edition*. *Cheshire, CT: Graphics Press*.

van Oort, F., de Geus, S., & Dogaru, T. (2015). Related Variety and Regional Economic Growth in a Cross-Section of European Urban Regions. *European Planning Studies*, *23*(6), 1110–1127. https://doi.org/10.1080/09654313.2014.905003

Ward, M., Grinstein, G., & Keim, D. (2010). *Interactive Data Visualization*.

Wickham, H. (2014). Tidy Data. *Journal of Statistical Software*, *59*(10). https://doi.org/10.18637/jss.v059.i10

Wickham, H. (2017). ggplot2 – Elegant Graphics for Data Analysis. *Journal of Statistical Software*, *77*(April), 3–5. https://doi.org/10.18637/jss.v077.b02

Yi, J. S., Kang, Y., Stasko, J. T., & Jacko, J. A. (2007). Toward a Deeper Understanding of the Role of Interaction in Information Visualization, *13*(6), 1224–1231.

Websites

EU Open Data portal
http://data.europa.eu/euodp/en/data/

ESIF Data
https://cohesiondata.ec.europa.eu/

European Data Portal
https://www.europeandataportal.eu/en/

Regional policy evaluation
http://ec.europa.eu/regional_policy/en/policy/evaluations/data-for-research/

Eurostat
http://ec.europa.eu/eurostat/web/json-and-unicode-web-services

Digital Agenda
https://ec.europa.eu/digital-single-market/en/policies/big-data

https://ec.europa.eu/digital-single-market/en/open-data

11 Annexes

11.1 Domains and IPCs

Table 18 – Semi-automated matching between technological domains and IPCs

Region	Needs/Challenges	Regional Domains	IPC 3 digits
Basilicata (5)	• Infrastructure gap and low internationalization • Low private investment and participation in the innovation system • Development of clusters and networks • Health and wellness improvement	Aerospace	A01,A23,B01,B25,B33 B64,B82,C12,G01,G02 G06,H01,H04
		Automotive	
		Green Economy	
		Energy	
		Tourism and culture industry	
Calabria (7)	• Reduction of the innovation gap and international development of SMEs • Improve Mobility and accessibility • Environmental sustainability • Health and wellness	Agrifood	A01,A23,A61,B01,B09,B65 C02,C04,C10,E02,E04,E05 F03,G01,G06,G08,G09
		Building / Green building	
		Tourism and culture	
		Environment and natural hazards	
		Life science	
		Logistics	
		ICT	
Campania (6)	• Development of smart cities and smart communities (mobility, health, education, energy, environment)	Advanced material and nanotechnologies	A01,A23,A61,A62,B09, B23,B28,B29,B32,B60,B61 B64,B82,C01,C02,C08,C09 C12,C21,D03,E04,F02,F15 F16,G01,G05,G06,G08 G09,G11,H01,H02,H04 Y02
		Aerospace	
		Energy, environment and green chemistry	
		Health biotechnologies and agrifood	
		Technologies for smart communities, cultural heritage, tourism and sustainable construction	
		Transport and advanced logistics	
Emilia-Romagna (5)	• Strengthening specialisation of the industrial system • Strengthening high-potential industrial systems • Development of Smart cities and communities • Improve Health and wellness • Development of innovation in the third sector	Mechatronics and automotive	A01,A21,A23,A61,B05 B25,B60,B65,D01,E04 E21,G06,H01,H02,H04
		Agrifood	
		Building	
		Cultural and creative industries	
		Health and wellness	

Region	Needs/Challenges	Regional Domains	IPC 3 digits
Friuli Venezia Giulia (6)	• Traditional manufacturing decline • Ageing society • Climate change • Energy vulnerability	Agrifood Home system Mechatronics Chemical and pharmaceutical Shipbuilding Cultural and creative industries	A01,A22,A23,A47,A61,B09 B23,B63,B82,C02,C12,E04 G01,G05,G06,H04
Lazio (7)	• Smart, Green and Integrated Transport • Health, Demographic change and wellness • Inclusive, innovative and reflective societies • Restoring, preserving, valuing & managing cultural heritage • Food security, sustainable agriculture and forestry, water research, bioeconomy • Secure, clean and efficient energy • Secure society	Aerospace Agrifood Creative digital industries Cultural patrimony and cultural technology Green economy Life science Security	A01,A23,A61,A62,A63 B29,B61,B64,E04,F02 G01,G06,G08,G09,G11 H04,H05
Liguria (3)	• Ageing society • Sustainability and development of the blue economy • Smart and Secure society • Climate change and environmental risk	Health and life science Marine technology Safety and quality of life	A01,A61,A62,B01,B09 B25,B60,B63,B82 C09,C12,E02,E04 G01,G06,H02
Lombardia (7)	• Ageing society • Health industry and wellness • Strengthening specialisation of industry and services • Environmental sustainability • Digital divide and smart society • Improve mobility and accessibility	Aerospace Agrifood Green manufacturing Health Artistic and cultural industries Advanced manufacturing Sustainable mobility	A23,A61,A62,A99,B01 B44,B60,B63,B64,B81 B82,C22,D01,G02,G06 G09,G21,H01,H02,H04 Y02
Marche (4)	• Ageing society • Traditional manufacturing decline • International competition • Brain-drain risk • Environmental risk	Domotics Mechatronics Health and Wellness Sustainable manufacturing	A23,A61,B07,B09,B25 B81,E04,G06,G08,H04
Molise (5)	• Digital divide and smart society	Buildings and smart cities	A23,A61,C02,E04,G06

Region	Needs/Challenges	Regional Domains	IPC 3 digits
	• Improve mobility and accessibility • Brain-drain risk	Automotive Agrifood ICT Life science	
Piemonte (5)	• Traditional manufacturing decline • Health and wellness improvement • Demographic change	Aerospace Automotive Made in Piemonte[50] Mechatronics Green Chemistry	A01,A23,A41,A61,B09,B60 B64,B82,C08,D01,G01,G05 G06,G07,G08,Y02
Provincia Autonoma di Bolzano (4)	• Moderate mountain depopulation • Marginal position in global production chain • Polarization of human resources specialisation • Low exploitation of cultural and linguistic plurality • Development of international collaboration • Environmental and quality of life improvement	Green Alpine Food Automation	A01,A23,A41,A61,B82 E04,G05,G06,H04
Provincia Autonoma di Trento (4)	• Improvement of population health and wellness • Slow technological transfer • Brain-drain risk • Development of traditional sectors	Agrifood Energy and environment Mechatronics Quality of life	A01,A21,A23,A61,A62 B09,B25,B33,B81,B82 C09,C10,C12,E04,E21 F01,F03,F16,G05,G06 G09,H02
Puglia (3)	• Ageing, disability and wellness • Gap with smart and digital communities (e.g. cities, tourism) • Low product innovation and competitiveness • Environmental and socially sustainable development	Smart communities[51] Health and environment[52] Sustainable manufacturing[53]	A23,A61,B09,B64,B81,B82 C08,D01,F03,F16,G06
Sardegna (5)	• Slowdown of ICT sector • Instability and low quality of energy supply • Support the agrifood dynamism • Support growth of the aerospace sector • Boost development of biomed technologies	Aerospace Agrifood Biomedicine Energy and environment ICT	A01,A21,A61,B63,B64,G01 G06,H02,Y02
Sicilia (6)	• Development of Smart cities and communities	Blue economy Agrifood	A01,A22,A23,A61,B63 C02,F03,G01,G06,H01

[50] Mainly agrifood and textile
[51] Cultural industries, social innovation, design and non-R&D innovation
[52] Wellness, Green and Blu economy, agrifood, tourism
[53] Smart industry, aerospace, mechatronics

Region	Needs/Challenges	Regional Domains	IPC 3 digits
	• Development of innovative products and services for inclusive and sustainable wellness • Strengthening the innovation capacity of specialisation sectors	Tourism and cultural heritage Smart cities and communities Energy Life Science	
Toscana (3)	• Fragmented productive sector • Low international development and technological transfer • Energy and environmental sustainability • Development of smart communities (mobility, health, energy, environment)	Chemistry and Nanotechnology ICT and photonics Smart manufacturing	A01,A61,B82 C02,F03,G06
Umbria (5)	• Development of the research-industry cooperation • Diversification of the productive environment • Development of Smart cities and communities	Agrifood Green chemistry Energy Smart manufacturing Life science	A01,A23,A61,B64,C08 C09,C12,E04,F03,G06
Valle d'Aosta (3)	• Stop the deindustrialization process • Accelerate diffusion, acquisition and development of innovation • Increase internationalization of the local economy • Increase human resources skills	Smart Mountain Excellent Mountain Green Mountain	A63,B09,C02,C10 E02,E04,G06,H02,H04
Veneto (4)	• Fragmented regional innovation system • Businesses and research gap • International competition • Low availability of relevant innovation and skills • Improve Mobility and accessibility	Smart Agrifood Creative industries Smart manufacturing Sustainable living	A01,A23,A41,A61 B60,B63,B65,C05,C10 E04,F25,G05,G06,Y02

11.2 Discussants assessment

VALUTAZIONE DELLA TESI DI DOTTORATO

Risposta all'indagine 1

ID risposta
683

Data invio
2019-01-04 13:46:02

Ultima pagina
9

Lingua iniziale
it

Seme
830512247

Partecipante
HRbKTAvQZW5mqIO

Data di inizio
2019-01-02 10:34:27

Data dell'ultima azione
2019-01-04 13:46:02

URL di riferimento

VALUTAZIONE DELLA TESI DI DOTTORATO

Titolo della tesi
Open Data Analytics - Advanced methods, tools and visualizations for policy making

Candidato
Roberto Palloni

Coordinatore del Corso di Dottorato
Prof. Francesco Piazza

Tutor/s Accademico
Emanuele Frontoni, Donato Iacobucci

Nome e affiliazione del valutatore
Alessandro Aldini, Università di Urbino Carlo Bo

VALUTAZIONE DELLA TESI DI DOTTORATO

Chiarezza ed organizzazione della tesi [Chiarezza]
Ottimo

Chiarezza ed organizzazione della tesi [Struttura della tesi]
Eccellente

Definizione degli obiettivi della ricerca [Obiettivi della ricerca]
Ottimo

Commento:
La tesi affronta il tema della data science in un contesto open data, discutendo e proponendo metodologie per la elaborazione e visualizzazione di informazioni a supporto dei processi decisionali nell'ambito specifico della pubblica amministrazione, con particolare riferimento al monitoraggio degli investimenti relativi ai fondi della Comunità Europea. La gestione ed analisi di big data, il principio dei dati aperti e le relative metodologie di rappresentazione e fruizione sono argomenti che hanno ricevuto ampio interesse in letteratura, e la stessa Commissione Europea ne stabilisce il ruolo critico nel valutare la qualità della propria amministrazione e nel supportare i processi decisionali. La prima parte della tesi tratta le strutture, le policy e le metodologie proposte in letteratura per l'utilizzo di open data a livello Europeo, sottolineando criticità e debolezze, tra le quali spicca ad esempio la usabilità. L'analisi è principalmente mirata a considerare l'efficacia e l'efficienza di monitoraggio e determinazione delle performance, a partire dalla specifica delle metriche di riferimento e dai livelli di aggregazione dei dati. La capacità di rendere disponibili i dati favorendone l'interpretazione è un elemento chiave, spesso sottovalutato, che rende necessario lo sviluppo di tecniche di visualizzazione dei dati che avanzino significativamente lo stato dell'arte. Un esempio è dato dal portale open data per ESIF, caratterizzato da livelli di aggregazione dei dati piuttosto limitanti. Le tecniche di visualizzazione devono rispondere a requisiti ben precisi per massimizzarne l'efficacia e possono impiegare approcci interattivi per ottenere questo risultato. L'analisi di tali requisiti, esposta nel capitolo 3, è utile a fornire le motivazioni e gli obiettivi che stanno alla base della metodologia e degli strumenti presentati successivamente, il cui principale risultato proposto è dato dal web tool ESIFY. Il concept del tool è chiaro così come i suoi obiettivi: offrire uno strumento interattivo di analisi di large open data che sia semplice, immediato, basato su un ampio set di metriche e in grado di estrapolare informazioni a qualunque livello di dettaglio disponibile. Il capitolo 5 offre un esempio di come l'analisi di open data guidata da indicatori formalizzati opportunamente (rispetto al dominio di interesse) possa fornire interpretazioni utili a supportare i processi strategici di valutazione e scelta, sebbene questo contributo possa essere messo in relazione più stretta e lineare con i risultati presentati nel capitolo precedente, specialmente in merito agli aspetti di semplicità, accessibilità ed immediatezza.

VALUTAZIONE DELLA TESI DI DOTTORATO

Grado di originalità della tesi [Originalità]
Ottimo

Commento:
Il carattere di originalità del principale contributo della tesi, espresso nei capitoli 4 e 5, è determinato e supportato dalla overview fornita nei capitoli precedenti, che mette in evidenza criticità e relative motivazioni a supporto degli approcci originali proposti. ESIFY è uno strumento che, come esposto in maniera chiara, fornisce quella flessibilità necessaria a superare i limiti noti di usabilità precedentemente messi in evidenza. L'analisi dei dati fornita nel capitolo 5 risulta in una tesi che fornisce una chiave di lettura interessante rispetto agli indicatori di monitoraggio relativi al dominio di interesse, sebbene il benchmark possa essere considerato non del tutto eterogeneo e limitato dalla mancanza di confronti su scala europea.

VALUTAZIONE DELLA TESI DI DOTTORATO

Adeguatezza e rigore metodologico [Adeguatezza della Metodologia]
Eccellente

Adeguatezza e rigore metodologico [Rigore metodologico]
Ottimo

Commento:
La metodologia presentata nel capitolo 4 è adeguata dal punto di vista tecnico e delle scelte progettuali fatte, e questo può essere facilmente evinto dalla chiara esposizione degli obiettivi, dell'approccio usato e dei risultati, motivati in maniera eccellente nei capitoli di overview. La formalizzazione della analisi proposta nel capitolo 5 è rigorosa ed i risultati che se ne deducono corretti e completi rispetto al benchmark considerato, e meritano di essere ulteriormente investigati in lavori futuri per stabilirne la variabilità su larga scala.

VALUTAZIONE DELLA TESI DI DOTTORATO

Risultati e correttezza delle conclusioni [Descrizione Risultati]
Eccellente

Risultati e correttezza delle conclusioni [Correttezza delle conclusioni]
Ottimo

Commento:
Le conclusioni riassumono il lavoro di tesi in maniera molto chiara offrendo una analisi che va oltre la semplice rassegna di quanto è stato presentato nei capitoli precedenti. Il candidato dimostra un eccellente spirito critico nel descrivere i propri risultati, mettendo in evidenza alcune tesi che si possono avanzare come risultato finale del lavoro svolto. La prima tesi (H1) è una conclusione piuttosto naturale del lavoro di rassegna svolto e giustifica, in un certo senso, l'approccio seguito per lo sviluppo del web tool ESIFY ed il relativo valore aggiunto rispetto allo stato dell'arte. E' condivisibile che il tool si presti a dimostrare in maniera semplice, anche agli occhi del non esperto di data analytics, fatti (parzialmente) noti che però richiedono evidenza a livello di interpretazione dei dati, mentre è più articolata e complessa la capacità di identificare relazioni del tutto sconosciute. Comunque il tool sposa in pieno il concetto di smart data fornendo lo strumento utile a rendere vivi i dati in tal senso. La seconda tesi (H2) è il risultato di uno studio analitico che sembra confermare il ruolo di alcuni indicatori nelle scelte strategiche di un dominio applicativo ben specifico. Le conclusioni su scala locale sono piuttosto interessanti e mettono in evidenza una realtà assolutamente non scontata o del tutto prevedibile, e pongono le basi per una futura analisi su larga scala al fine di chiarire ulteriormente aspetti che possono essere decisivi per il prossimo programma quadro settennale della Commissione Europea.

VALUTAZIONE DELLA TESI DI DOTTORATO

Contributo al progresso della conoscenza nel settore [Contributo alla conoscenza]
Ottimo

Commento:
La tesi fornisce diversi nuovi contributi allo stato dell'arte. In primo luogo mette a disposizione una overview esaustiva della letteratura che evidenzia requisiti, criticità e problematiche degli attuali sistemi di data analytics per open data nel contesto delle pubbliche amministrazioni con particolare rilievo per la Comunità Europea. In particolare, l'analisi puntuale degli strumenti attualmente a disposizione giustifica la necessità di sviluppare uno strumento capace di coniugare semplicità (nella interpretazione dei dati) e flessibilità (nel considerare aggregazioni a diversi livelli di granularità). Il risultato e contributo principale in tal senso è i lweb tool ESIFY, che può rappresentare uno strumento molto utile a supporto di tutti i processi di valutazione e decisionali. Un ulteriore contributo della tesi è dato dall'approccio metodologico seguito per l'analisi descritta nel capitolo 5, che non solo risulta in conclusioni originali, ma pone le basi per ulteriori indagini condotte con lo stesso rigore formale.

VALUTAZIONE DELLA TESI DI DOTTORATO

Potenziale impatto della ricerca e applicabilità dei risultati [Impatto]
Eccellente

Potenziale impatto della ricerca e applicabilità dei risultati [Applicabilità]
Eccellente

Commento:
I risultati esposti nella tesi hanno una immediata e fruibile applicazione in contesti e domini applicativi reali, essendo ottenuti a partire da situazioni ed esigenze estremamente concrete (come può infatti essere il problema dell'open data analytics a supporto delle attività di valutazione e decisionali nel mondo ESIF per la Comunità Europea) attraverso però lo sviluppo di metodologie e strumenti in modo rigoroso e formale, aspetto che pone le basi per garantire un impatto su più ampia scala (ad esempio, non limitatamente alle analisi condotte nel contesto della Comunità Europea). Anche i risultati ottenuti dallo studio analitico relativo ad un dominio specifico e su un benchmark limitato, oltre ad offrire risultati specifici di per sé interessanti ed immediatamente fruibili ed usabili, offrono nuovi spunti (anche in ambito metodologico) che possono trovare immediata applicabilità in domini applicativi analoghi.

VALUTAZIONE DELLA TESI DI DOTTORATO

Adeguatezza / Completezza dei riferimenti [Riferimenti]
Eccellente

Commento:
La bibliografia fornita nella tesi è ampia, adeguata e completa, coprendo non solo riferimenti alla letteratura scientifica, ma anche i necessari riferimenti comunitari (legislativi, di policy definition e data assessment) che rendono la tesi uno strumento di consultazione ad ampio spettro per i diversi stakeholders interessati all'argomento dell'open data analytics in un contesto Europeo. La bibliografia è completa anche in merito alle fonti dei dati e dei benchmark utilizzati all'interno della tesi.

Altre note e commenti

VALUTAZIONE COMPLESSIVA

La tesi soddisfa i requisiti minimi per essere ammessa all'esame finale ?
Si

In caso di valutazione positiva e ammissione all'esame finale, specificare
La tesi può essere presentata per l'esame finale

In caso di valutazione positiva e ammissione all'esame finale la valutazione complessiva è: [Valutazione complessiva]
Ottimo

in caso di necessarie ulteriori revisioni sostanziali i seguenti punti dovrebbero essere corretti / migliorati / integrati: (fornire commenti e suggerimenti su ogni aspetto che dovrebbe essere migliorato)

Data di compilazione
2019-01-04 00:00:00

VALUTAZIONE DELLA TESI DI DOTTORATO

Risposta all'indagine 1

ID risposta
986

Data invio
2019-01-16 10:41:36

Ultima pagina
9

Lingua iniziale
it

Seme
1363856582

Partecipante
FzO5ekJ4rsz8p6a

Data di inizio
2019-01-15 17:40:03

Data dell'ultima azione
2019-01-16 10:41:36

URL di riferimento

Nome
Fratesi Ugo

Cognome
Prof./Prof.ssa

Indirizzo e-mail
ugo.fratesi@polimi.it

corsodottorato
INGEGNERIA DELL'INFORMAZIONE

coordinatore
Prof. Francesco Piazza

emailcoordinatore
f.piazza@univpm.it

dottorando
Palloni Roberto

emaildottorando
r.palloni@pm.univpm.it

VALUTAZIONE DELLA TESI DI DOTTORATO

Titolo della tesi
Open Data Analytics Advanced methods, tools and visualizations for policy making

Candidato
Roberto Palloni

Coordinatore del Corso di Dottorato
Prof. Francesco Piazza

Tutor/s Accademico
Emanuele Frontoni Donato Iacobucci

Nome e affiliazione del valutatore
Ugo Fratesi, Politecnico di Milano

VALUTAZIONE DELLA TESI DI DOTTORATO

Chiarezza ed organizzazione della tesi [Chiarezza]
Eccellente

Chiarezza ed organizzazione della tesi [Struttura della tesi]
Ottimo

Definizione degli obiettivi della ricerca [Obiettivi della ricerca]
Eccellente

Commento:
Gli obiettivi della tesi sono ben chiari fin dal principio, anche se poi, nel testo, nonostante la struttura numerata, il filo organizzativo è perfettibile (si vedano i piccoli suggerimenti).

VALUTAZIONE DELLA TESI DI DOTTORATO

Grado di originalità della tesi [Originalità]
Eccellente

Commento:
La tesi a mio avviso è molto originale, in quanto è un primo tentativo di collegare un approccio di sistemi informativi all'analisi delle politiche di smart specialisation.

VALUTAZIONE DELLA TESI DI DOTTORATO

Adeguatezza e rigore metodologico [Adeguatezza della Metodologia]
Eccellente

Adeguatezza e rigore metodologico [Rigore metodologico]
Eccellente

Commento:
Non mi posso esprimere sulla prima parte in quanto afferente a un'altra disciplina scientifica. Nella seconda parte il rigore metodologico è ottimo e la metodologia adeguata. In particolare, ho apprezzato il fatto che in molti casi la tesi discuta i limiti degli indicatori che propone e delle fonti statistiche disponibili

VALUTAZIONE DELLA TESI DI DOTTORATO

Risultati e correttezza delle conclusioni [Descrizione Risultati]
Buono

Risultati e correttezza delle conclusioni [Correttezza delle conclusioni]
Ottimo

Commento:
A mio avviso i risultati della seconda parte (non posso commentare quelli della prima) sono molto interessanti e proprio per questo potrebbero essere valorizzati meglio, messi più in evidenza e costruire un percorso a partire da ipotesi ex-ante più definite. La costruzione sistematica di indicatori in grado di valutare ex-ante le politiche di smart specialisation è già un risultato notevole, in particolare quando essi sono collegati all'analisi teorica, ma anche i risultati sono interessanti e meritano di essere evidenziati meglio.

VALUTAZIONE DELLA TESI DI DOTTORATO

Contributo al progresso della conoscenza nel settore [Contributo alla conoscenza]
Ottimo

Commento:
Ritengo che i risultati ottenuti siano molto interessanti e pertanto abbiano il potenziale di essere valorizzati in papers scientifici da pubblicare su riviste internazionali

VALUTAZIONE DELLA TESI DI DOTTORATO

Potenziale impatto della ricerca e applicabilità dei risultati [Impatto]
Eccellente

Potenziale impatto della ricerca e applicabilità dei risultati [Applicabilità]
Eccellente

Commento:
La ricerca è sicuramente interessante e presenta due contributi di chiara applicabilità con un importante impatto. Da un lato, la visualizzazione dei dati delle politiche di coesione in modo più facilmente interpretabile ed utilizzabile, così da fornire un supporto alle decisioni di policy. Dall'altro, l'analisi delle strategie di smart specialisation effettuata nella tesi, sperimentalmente sull'Italia, potrebbe costituire un importante contributo alla verifica della rispondenza delle suddette politiche al quadro teorico di riferimento, al fine di migliorare la programmazione regionale.

VALUTAZIONE DELLA TESI DI DOTTORATO

Adeguatezza / Completezza dei riferimenti [Riferimenti]
Ottimo

Commento:
Commento anche qui esclusivamente sulla seconda parte. I riferimenti alla letteratura sulla smart specialisation appaiono adeguati e aggiornati, anche se sarebbe stato preferibile presentarli come capitolo a parte anziché capitolo 5.1. Ho apprezzato in particolare che gli indicatori vengano costruiti sulla base della teoria e vadano a seguire i concetti teoricamente delineati.

Altre note e commenti
Consiglio di: Spostare le ipotesi formulate dalle conclusioni all'introduzione, per valorizzarle meglio Espandere la discussione dei risultati sulle strategie di smart specialisation italiane, in quanto ci sono già molti elementi di potenziale interesse per i policy makers che emergono dall'analisi Rendere chiaro il passaggio che si fa a pagina 93, dove attualmente c'è un salto da un'analisi per l'Italia a un'analisi per l'Europa basata sulle collaborazioni Cordis. Consiglio di inserirvi un titolo e un'introduzione che faccia capire il senso della nuova analisi e il suo legame con l'analisi precedente. Strutturare di più il capitolo 5, che si potrebbe dividere, gerarchicamente, in almeno tre capitoli, uno di literature review, uno di indicatori e uno di risultati. Fare attenzione (pagina 78, per esempio) al fatto che la strategia di smart specialisation non solo dovrebbe identificare i domini tecnologici presenti, ma quelli con potenzialità nella regione. Pertanto, anche se si riscontra generalmente un certo ottimismo dei policy makers nell'identificare le potenzialità della loro regione, non possiamo considerare l'indicatore di coerenza costruito come un valore assoluto. A pagina 80 (figura 36 e nota 41) mi chiedo se c'è anche un effetto puramente statistico dietro il risultato, in quanto maggiore è il numero di IPC scelti, maggiore è, probabilisticamente, il numero di IPC scelti tra quelli presenti nella regione. Fare attenzione nella scrittura a rendere sempre esplicito e chiaro il livello spaziale (NUTS) utilizzato nelle varie analisi

VALUTAZIONE COMPLESSIVA

La tesi soddisfa i requisiti minimi per essere ammessa all'esame finale ?
Si

In caso di valutazione positiva e ammissione all'esame finale, specificare
La tesi ha bisogno di revisioni non sostanziali da produrre entro 30 giorni

In caso di valutazione positiva e ammissione all'esame finale la valutazione complessiva è: [Valutazione complessiva]
Ottimo

in caso di necessarie ulteriori revisioni sostanziali i seguenti punti dovrebbero essere corretti / migliorati / integrati: (fornire commenti e suggerimenti su ogni aspetto che dovrebbe essere migliorato)

Data di compilazione
2019-01-15 00:00:00